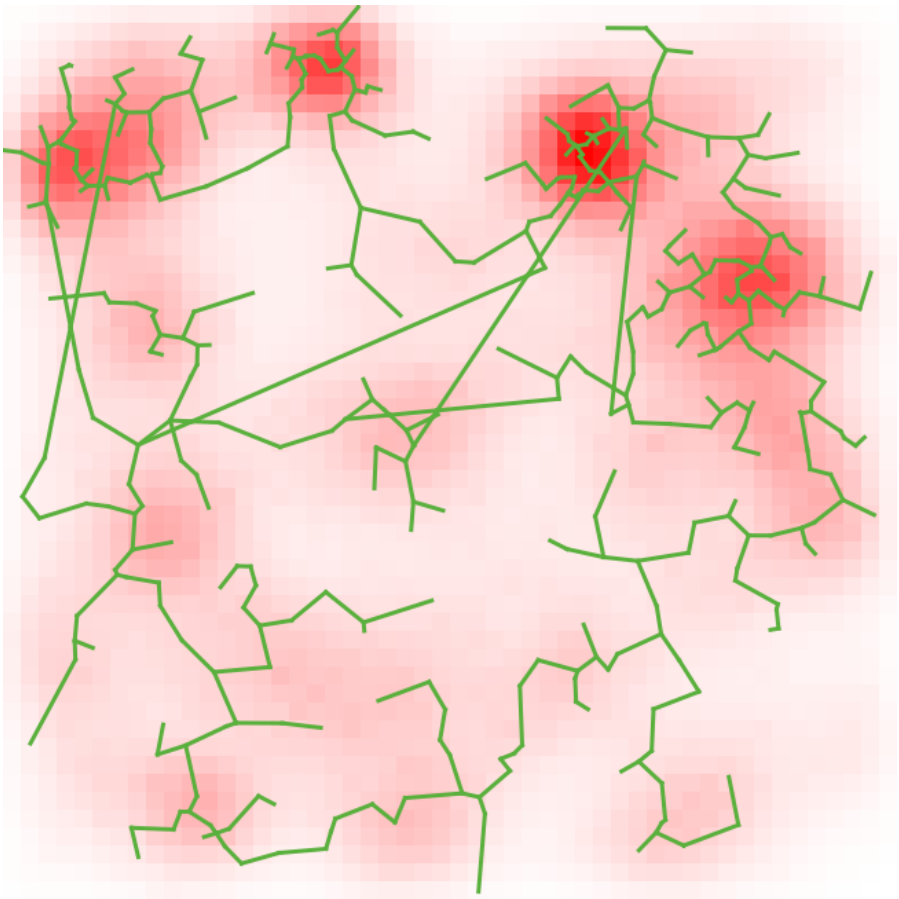
VERS DES MODÈLES COUPLANT DÉVELOPPEMENT URBAIN ET CROISSANCE DES RÉSEAUX DE TRANSPORT

JUSTE RAIMBAULT



Mémoire de Thèse de Doctorat

Under the supervision of Arnaud Banos and Florent Le Néchet

UMR CNRS 8504 Géographie-cités
and UMR-T IFSTTAR 9403 LVMT

Université Paris Diderot - Paris 7

July 2017 – version 3.2

ABSTRACT

Résumé

C : (Florent) trop de concepts dans l'abstract, peut pas apporter qqchse à tous

C : (Florent) commencer par expliquer ce que sont causalités circulaires et pourquoi difficiles à modéliser

C : (Arnaud) complexly : ?

C : (Arnaud) théorie des systèmes territoriaux en réseau co-évolutifs ?

NOTES DE LECTURE

PUBLICATIONS

Les travaux suivants contiennent une grande partie du contenu de cette thèse :

PUBLICATIONS

Antelope, C., Hubatsch, L., Raimbault, J., and Serna, J. M. (2016). An interdisciplinary approach to morphogenesis. Forthcoming in Proceedings of Santa Fe Institute CSSS 2016.

Raimbault, J. (2017). A Discrepancy-Based Framework to Compare Robustness Between Multi-attribute Evaluations. In Complex Systems Design & Management (pp. 141-154). Springer International Publishing. [Raimbault, 2016a]

Raimbault, J. (2016). Investigating the Empirical Existence of Static User Equilibrium, *forthcoming in EWGT 2016 proceedings, Transportation Research Procedia.* arxiv :1608.05266 [Raimbault, 2016d]

Raimbault, J. (2016). Generation of Correlated Synthetic Data, forthcoming in *Actes des Journées de Rochebrune 2016.*

Raimbault, J. (2015). Models Coupling Urban Growth and Transportation Network Growth : An Algorithmic Systematic Review Approach, forthcoming in *ECTQG 2015 proceedings.* arxiv :1605.08888

COMMUNICATIONS

Towards a Theory of Co-evolutive Networked Territorial Systems : Insights from Transportation Governance Modeling in Pearl River Delta, China, *MEDIUM Seminar : Sustainable Development in Zhuhai, Guangzhou, Dec 2016.*

Models of growth for system of cities : Back to the simple, *Conference on Complex Systems 2016, Amsterdam, Sep 2016.*

For a Cautious Use of Big Data and Computation. *Royal Geographical Society - Annual Conference 2016 - Session : Geocomputation, the Next 20 Years (1), London, Aug 2016.*

Indirect Bibliometrics by Complex Network Analysis. *20e Anniversaire de Cybergeo, Paris, May 2016.*

Raimbault, J. & Serra, H. (2016). Game-based Tools as Media to Transmit Freshwater Ecology Concepts, *poster corner at SETAC 2016 (Nantes, May 2016).*

Le Néchet, F. & Raimbault, J. (2015). Modeling the emergence of metropolitan transport authority in a polycentric urban region, *ECTQG 2015, Bari, Sep 2015).*

Hybrid Modeling of a Bike-Sharing Transportation System, *poster presented at ICCSS 2015, Helsinki, June 2015.*

Raimbault, J. & Gonzales, J. (2015). Application de la Morphogénèse de Réseaux Biologiques à la Conception Optimale d'Infrastructures de Transport, *poster presented at Rencontres du Labex Dynamite, Paris, May 2015.*

ACKNOWLEDGEMENTS

Un certain nombre de résultats obtenus dans cette thèse ont été calculés sur l'organisation virtuelle vo.complex-system.eu de l'European Grid Infrastructure (http ://www.egi.eu). Nous remercions l'European Grid Infrastructure et ses National Grid Initiatives (France-Grilles en particulier) pour fournir le support technique et l'infrastructure.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

C : (Florent) cf receuil articles du Monde sur Grd Paris (numériser)

C : (Florent) HDR Anne ?

C : (Florent) trop peu ancré concrètement dans le champ des interactions transport/ville - enchainement idée ok mais revoir granularité info. Catalogue de situations complexes d'interactions forme urbaine/transport à reproduire.

INTRODUCTION

INTRODUCTION

C'est quand on donne un coup de pied dans la fourmilière qu'on se rend compte de toute sa complexité.
- Arnaud Banos

C : (Florent) cet exemple parait loin de l'approche ? il n'est pas territorial, bien mais HS ; un autre sur dynamiques de certaines villes connectées ou non serait plus approprié

"En conséquence d'un problème technique, le trafic est interrompu sur la ligne B du RER pour une durée indéterminée. Plus d'information seront fournies dès que possible". Il y a des fortes chances pour que quiconque ayant vécu ou passé un peu de temps en région parisienne ait déjà entendu cette annonce glaçante et en ait subi les conséquences pour le reste de la journée. Mais il ne se doute sûrement pas des ramifications des cascades causales induites par cet évènement presque banal. Les systèmes territoriaux, quelles que soient les aspects considérés pour leur définition, seront toujours extrêmement complexes, les interrelations à de nombreuses échelles spatiales et temporelles participant à la production des comportements émergents observés à tout niveau du système. Martin est un étudiant qui fait l'aller-retour journalier entre Paris et Palaiseau and manquera un examen crucial, ce qui aura un impact profond sur sa vie professionnelle : implications à une longue échelle de temps, une petite échelle spatiale et à la granularité de l'agent. **C :** (Florent) ? Yuangsi était en train de relier les aéroports d'Orly et Roissy dans son voyage de Londres à Pékin et va manquer son avion ainsi que le mariage de sa soeur : grande échelle spatiale, petite échelle de temps, granularité de l'agent. Une pétition collective émerge des voyageurs, conduisant à la création d'une organisation qui mettra la pression sur les autorités pour qu'elles augmentent le niveau de service : échelle temporelle et spatiales mesoscopique, granularité de l'aggregation d'agents. La recherche de cause possible à l'incident conduira à des processus intriqués à diverses échelles, parmi lesquels aucun ne semble être une meilleure explication ; le développement historique du réseau ferroviaire en région parisienne a conditionné les écolutions futures et le RER B a suivi l'ancienne Ligne de Sceaux, le plan de Delouvrier pour le développement régional et son execution partielle, sont également des éléments d'explication des faiblesses structurelles du réseau parisien de transports en commun [Gleyze, 2005] **C :** (Florent) réseau parisien un des plus résilients du monde, cf slides Erik Janus

KTH ; les motifs pendulaires dus à l'organisation territoriale induisent une surcharge de certaines ligne et ainsi nécessairement une augmentation des incidents d'exploitation. La liste pourrait être ainsi continuée un certain temps, chaque approche apportant sa vision mature correspondant à un corpus de connaissances scientifiques dans des disciplines diverses comme la géographie, l'économie urbains, les transports. Cette anecdote amusante est suffisante pour faire ressentir la complexité des systèmes territoriaux. Notre but ici est de se plonger dans cette complexité, et en particulier donner un point de vue original sur l'étude des relations entre réseaux et territoires. Le choix de cette position sera largement discuté dans une partie thématique, nous nous concentrons à présent sur l'originalité du point de vue que nous allons prendre.

DE LA POSITION GÉNÉRALE

L'ambition de cette thèse est de ne pas avoir d'ambition. Cette entrée en matière, rude en apparence, contient à différents niveaux les logiques sous-jacentes à notre processus de recherche. Au sens propre, nous nous plaçons tant que possible dans une démarche constructive et exploratoire, autant sur les plans théoriques et méthodologiques que thématique, mais encore proto-méthodologique (outils appliquant la méthode) : si des ambitions unidimensionnelles ou intégrées devaient émerger, elles seraient conditionnées par l'arbitraire choix d'un échantillon temporel parmi la continuité de la dynamique qui structure tout projet de recherche. Au sens structurel, l'auto-référence qui soulève une contradiction apparente met en exergue l'aspect central de la réflexivité dans notre démarche constructive, autant au sens de la récursivité des appareils théoriques, de celui de l'application des outils et méthodes développés au travail lui-même ou que de celui de la co-construction des différentes approches et des différents axes thématiques. Le processus de production de connaissance pourra ainsi être lu comme une métaphore des processus étudiés. Enfin, sur un plan plus enclin à l'interprétation, cela suggérera la volonté d'une position délicate liant un positionnement politique dont la nécessité est intrinsèque aux sciences humaines (par exemple ici contre l'application technocratique des modèles, ou pour le développement d'outils luttant pour une science ouverte) à une rigueur d'objectivité plus propre aux autres champs abordés, position forçant à une prudence accrue.

CONTEXTE SCIENTIFIQUE : PARADIGMES DE LA COMPLEXITÉ

Pour une meilleure introduction du sujet, il est nécessaire d'insister sur le cadre scientifique dans lequel nous nous positionnons. Ce contexte est crucial à la fois pour comprendre les concepts épistémo-

logiques implicites dans nos questions de recherche, et aussi pour être conscient de la variété de méthodes et outils utilisés. La science contemporaine prend progressivement le tournant de la complexité dans de nombreux champs **C :** (Florent) tout le monde ne connait pas , ce qui implique une mutation épistémologique pour abandonner le réductionnisme strict qui a échoué dans la majorité de ses tentatives de synthèse [Anderson, 1972]. Arthur a rappelé récemment [Arthur, 2015] qu'une mutation des méthodes et paradigmes en était également un enjeu, de par la place grandissante prise par les approches computationnelles qui remplacent les résolutions purement analytiques généralement limité en possibilités de modélisation et de résolution. La capture des *propriétés émergentes* par des modèles de systèmes complexes est une des façons d'interpréter la philosophie de ces approaches.

C : (Florent) rebondir sur thématique, questce qui emerge

Ces considérations sont bien connues des Sciences Humaines (qualitatives et quantitatives) pour lesquelles la complexité des agents et systèmes étudiés est une des justifications de leur existence : si les humains étaient des particules, la majorité des disciplines les prenant comme objet d'étude n'auraient jamais émergé puisque la thermodynamique aurait alors résolu la majorité des problèmes sociaux **C :** (Florent)attention phrases asimoviennes [1]. Elles sont au contraire moins connues et acceptées en sciences "dures" comme la physique : Laughlin développe dans [Laughlin, 2006] une vision de la discipline **C :** (Florent)which ? à la même position de "frontière des connaissances" que d'autre champs pouvant paraître moins matures. La plupart des connaissances actuelles concerne des structures classiques simples, alors qu'un grand nombre de système présentent des propriétés *d'auto-organisation*, au sense ou les lois macroscopiques ne sont pas suffisantes pour inférer les propriétés macroscopiques du systèmes à moins que son évolution soit entièrement simulée (plus précisément cette vision peut être prise comme une définition de l'émergence sur laquelle nous reviendrons par la suite, or des propriétés auto-organisées sont par nature émergentes). Cela correspond au premier cauchemar du Démon de Laplace développé dans [Defuant et al., 2015].

A la croisée de positionnements épistémologiques, de méthodes et de champs d'application, les *Sciences de la complexité* se concentrent sur l'importance de l'émergence et de l'auto-organisation dans la plupart des phénomènes réel, ce qui les place plus proche de la frontière des connaissances que ce que l'on peut penser pour des disciplines classiques (Laughlin, op. cit.). Ces concepts ne sont pas récents et avaient déjà été mis en valeur par Anderson [Anderson,

1 bien que cette affirmation soit elle-même discutable, les sciences physiques classiques ayant également échoué à prendre en compte l'irréversibilité et l'évolution de Systèmes Complexes Adaptatifs comme le souligne Prigogine dans [Prigogine et Stengers, 1997].

1972]. On peut aussi interpréter la Cybernétique comme un précurseur des Sciences de la Complexité en la lisant comme un pont entre technologie et sciences cognitives [Wiener, 1948]. **C :** (Florent) pquoi parler de ca ici ? Plus tard, la Synergétique [Haken, 1980] a posé les bases d'approches théoriques des phénomènes collectifs en physique. Les causes possibles de la croissance récente du nombre de travaux se réclamant d'approches complexes sont nombreuses. L'explosion de la puissance de calcul en est certainement une vu le rôle central que jouent les simulations numériques [Varenne, 2010b]. Elles peuvent aussi être à chercher auprès de progrès en épistémologie : introduction de la notion de perspectivisme [Giere, 2010c], reflexions plus fine autour de la nature des modèles [Varenne et Silberstein, 2013][2]. Les potentialités théoriques et empiriques de telles approchent jouent nécessairement un rôle dans leur succès[3], comme le confirme les domaines très variés d'application (voir [Newman, 2011] pour une revue très générale), comme par exemple la Science de Réseaux [Barabasi, 2002] ; les Neurosciences [Koch et Laurent, 1999] ; les Sciences Sociales ; la Géographie [Manson, 2001][Pumain, 1997] ; la Finance avec les approches écononophysiques [Stanley et al., 1999] ; l'Ecologie [Grimm et al., 2005]. La Feuille de Route des Systèmes Complexes [Bourgine, Chavalarias et al., 2009] propose une double lecture des travaux en Complexité : une approche horizontale faisant la connexion entre champs d'étude par des questions transversales sur les fondations théoriques de la complexité et des faits stylisés empiriques communs, et une approche verticale, dans le but de construire des disciplines intégrées et les modèles multi-scalaires hétérogènes correspondants. L'interdisciplinarité est ainsi cruciale pour notre contexte scientifique.

C : (Florent) donner ici exemples dans champ transports/urba

C : (Florent) plus de détails sur les disciplines CS ?

interdisciplinarité

Il est important d'insister sur le rôle de l'interdisciplinarité dans la position de recherche prise ici. Il s'agit moins d'un travail en Géographie ou en Modélisation de Systèmes Complexes Adaptatifs, pouvant difficilement être vraiment les deux à la fois, mais en *Science des Systèmes Complexes* que nous réclamons discipline propre comme le propose Paul Bourgine. **C (Florent) :** pas vraiment fondateur de la discipline **A1 :** non mais du point de vue particulier que nous défendons - théories intégratives roadmap etc. - trouver une ref là dessus ?

2 dans ce cadre, les progrès scientifiques et épistémologiques ne peuvent pas être dissociés et peuvent être vus comme étant en co-évolution

3 même si l'adoption de nouvelles pratiques scientifiques est souvent largement biaisé par l'imitation et le manque d'originalité [Dirk, 1999], ou de façon plus ambivalente, par des stratégies de positionnement puisque le combat pour les fonds est un obstacle croissant à une recherche saine [Bollen et al., 2014].

Ce n'est pas sans risques d'être lu avec méfiance voir défiance par les tenants des disciplines classiques, comme des exemples récents de malentendus ou conflits ont récemment illustré [Dupuy et Benguigui, 2015]. Il faut se rappeler l'importance de la spirale vertueuse de Banos entre disciplinarité et interdisciplinarité [Banos, 2013]. Celle-ci doit nécessairement impliquer différents agents scientifiques, et il est compliqué pour un agent de se positionner dans les deux branches ; notre fond scientifique ne nous permet pas de nous positionner dans la *disciplinarité géographique* mais bien dans celle des Systèmes Complexes (qui est interdisciplinaire, voir 3.3 pour contourner la contradiction apparente), et notre sensibilité scientifique et épistémologique nous pousse à faire de même.

Le positionnement de Batty lorsqu'il propose *Une Nouvelle Science des Villes* [Batty, 2013b] (qu'il présente avec humour comme *La* nouvelle science des villes), se présente comme une intégration des disciplines et méthodes vers une science définie par son objet d'étude, les villes. Its theoretical and epistemological weaknesses (no theoretical constructions of studied geographical objects on the one hand, approximative contextualization of complexity) combined with an overall impression of *pot-pourri* of forgotten works (space syntax, land-use models), unfortunately avoid us to use it as we will use geographical theories (e.g. evolutive urban theory) in an appropriated epistemological complexity context. Yet our reading of this work may be the result of a misunderstanding due to different cultural backgrounds. **C (Arnaud) :** j'espère que tu abuses ? :)!! Argument d'autorité **A1 :** yes, changer positionnement complètement malvenu **C (Florent) :** attention arguments autorité ; insister sur difficulté à intégrer paradigmes plutot que juger précédents **A1 :** idem

L'évolution scientifique des sciences de la complexité, qui est vue par certains comme une révolution [Colander, 2003], ou même comme *un nouveau type de science*, pourrait affronter des difficultés intrinsèques dues aux comportements et a-priori des chercheurs en tant qu'être humains. **C :** (Florent) idem développer transport/transports/modeling (?) Plus précisément, le besoin d'interdisciplinarité qui fait la force des Sciences de la Complexité pourrait devenir une de ses grandes faiblesses, puisque la structure fortement en silo de la science peut avoir des impacts négatifs sur les initiatives impliquant des disciplines variées. Nous n'évoquons pas les problèmes de sur-publication, quantification, competition, qui sont plus liés à des questions de Science Ouverte et de son éthique, tout aussi de grande importance mais d'une autre nature. Cette barrière qui nous hante et que nous pourrions ne pas surmonter, a pour plus évident symptôme des *divergences culturelles disciplinaires*, et les conflits d'opinion en résultant. Ce drame du malentendu scientifique est d'autant plus grave qu'il peut en effet détruire totalement certains progrès en interprétant comme une falsification des travaux qui traitent une ques-

tion toute différente. L'exemple récent d'un travail sur les inégalités liées aux hauts revenus présenté dans [Aghion et al., 2015], et dont les conclusions ont été commentées comme s'opposant aux thèses de Piketty dans [Piketty, 2013], est typique de ce schéma. Alors que Piketty se concentre sur la construction de bases de données propres sur le temps long pour les revenus et montre empiriquement une récente accélération des inégalités de revenus, son modèle visant à lier ce fait stylisé avec l'accumulation de capital a été critiqué comme sur-simplifié. D'autre part, Bergeaud *et al.* montrent par un modèle d'économie de l'innovation que *sous certaines hypothèses* les écarts de revenus peuvent être bénéfique à l'innovation et donc à une utilité globale. D'où des conclusions divergentes sur le rôles des capitaux personnels dans une économie. **C :** (Florent) hors-sujet, reste ds domaine (?) Mais des *point de vus* ou *interprétations* différentes ne signifient pas une incompatibilité scientifique, et on pourrait même imaginer rassembler ces deux approches dans un cadre et modèle unifié, produisant des interprétations possiblement similaires et potentiellement encore nouvelles. Une telle approche intégrée aura de grandes chances de contenir plus d'information (selon comment le couplage est opéré) et être une avancée scientifique. Cette expérience de pensée illustre les potentialités et la nécessité de l'interdisciplinarité. Dans une autre veine assez similaire, [Holmes et al., 2017] ré-analyse des données biologiques d'une expérience de 1943 qui prétendait confirmer l'hypothèse des processus d'évolution Darwiniens par rapport aux processus Lamarckiens, et montrent que les conclusions ne tiennent plus dans le contexte actuel d'analyse de données (avances énormes sur la théorie et les possibilités de traitement) et scientifique (avec d'autre nombreuses preuves de nos jours des processus Darwiniens) : c'est un bon exemple de malentendu sur le contexte, et comment le cadre de travail à la fois technique et thématique influence fortement les conclusions scientifiques. Nous développons à présent divers exemples révélateurs de la manière dont des conflits entre disciplines peuvent être dommageables.

la tentation de réinventer la géographie Comme déjà mentionné, Dupuy et Benguigui soulignent dans [Dupuy et Benguigui, 2015] le fait que les sciences urbaines **C :** (Florent) définition ? ont récemment connu des conflits ouverts entre les tenants classiques des disciplines et des nouveaux arrivants, en particulier les physiciens. **C :** (Florent) gravité de Wilson par max entropie n'est pas nouveau La disponibilité de grand jeux de données d'un nouveau type (réseaux sociaux, données des nouvelles technologies de la communication) ont attiré leur attention sur des objets plus traditionnellement étudiés par les sciences humaines, puisque les méthodes analytiques et computationnelles de la physique statistique sont devenues applicables. Bien que ces travaux soient généralement présentés

comme la construction d'une approche scientifique des villes, tout en impliquant que la connaissance existante n'est pas scientifique de par sa nature plus qualitative, ils n'ont aucunement révélé de connaissance nouvelle sur les systèmes urbains : **C :** (Florent) pas nécessaire dans la thèse pour citer quelques exemples, [Barthelemy et al., 2013] conclut que Paris a subit une transition pendant la période d'Haussman et ses opérations de planification globale, qui sont des faits naturellement connus depuis longtemps en Histoire Urbaine et Géographie Urbaine. [Chen, 2009] redécouvre que le modèle gravitaire est amélioré par l'introduction de décalages dans les interactions et dérive analytiquement l'expression d'une force d'interaction entre les villes, sans aucun cadre théorique ni thématique. De tels exemples peuvent être multipliés, confirmant l'inconfort courant entre physiciens et géographes. Des bénéfices significatifs pourraient résulter d'une intégration raisonnée des disciplines [O'Sullivan et Manson, 2015] mais la route semble être bien longue encore.

C : (Florent) a developper, conceretement, quels verrous à faire sauter ?

economie géographie ou géographie economique ? Des conflits similaires se rencontrent en économie : comme décrit par [Marchionni, 2004], la discipline de l'économie géographique, traditionnellement proche de la géographie, a fortement critiqué un nouveau courant de pensé nommé *économie géographisée*, **C :** (Arnaud) New economic geography ? dont le but est la spatialization des techniques économiques classiques. Chacune n'ont pas les mêmes desseins et buts, et le conflit apparaît comme un malentendu complet vu d'un oeil extérieur.

C : (Florent) a developper ou ne pas en parler, un peu loin du coeur du sujet tel que abordé

modélisation basée agent en economie Des conflits disciplinaires peuvent aussi se manifester sous la forme d'un rejet de méthodes nouvelles par les courants dominants. Suivant Farmer [Farmer et Foley, 2009], l'échec opérationnel de la plupart des approches économiques classiques pourrait être compensé par un usage plus systématique de la modélisation et simulation basées agent. L'absence de cadre analytique qui est naturelle pour l'étude de la plupart des systèmes complexes adaptatifs semble rebuter la plupart des économistes. **C :** (Florent) contraire sans doute vrai aussi

C : (Arnaud) Difficile de se positionner de manière crédible sur ces sujets en 5 lignes et 1 reference !

finance La finance quantitative peut être instructive pour notre propos et sujet, d'une part par les similarités de la cuisine interdisciplinaire avec notre domaine (rapport avec la physique et l'éco-

nomie, champs plus ou moins “rigoureux”, etc.). Dans ce domaine coexistent divers champs de recherche ayant très peu d’interactions entre eux. On peut considérer deux exemples. D’une part, les statistiques et l’économétrie sont extrêmement avancées en mathématiques théoriques, utilisant par exemple des méthodes de calcul stochastique et de théorie des probabilité pour obtenir des estimateurs très raffinés de paramètres pour un modèle donné (voir par exemple [Barndorff-Nielsen et al., 2011]). D’autre part, l’éconophysique a pour but d’étudier des faits stylisés empiriques et inférer les lois correspondantes pour tenter d’expliquer les phénomènes liés à la complexité des marchés financiers [Stanley et al., 1999], comme par exemple les cascades menant aux ruptures de marché, les propriétés fractales des signaux des actifs, la structure complexe des réseaux de corrélation. Chacun a ses avantages dans un contexte particulier et gagnerait à des interactions accrues entre les deux domaines.

Ces divers exemples pris au fil du vent sont de brèves illustrations du caractère crucial de l’interdisciplinarité et de sa difficulté à pratiquer. Sans presque exagérer, on pourrait imaginer l’ensemble des chercheurs se plaindre de mauvaises ou difficiles expériences d’interdisciplinarité, avec un retour largement positif lors des rares succès. Nous allons tenter par la suite d’emprunter ce chemin étroit, empruntant des idées, théories et méthodes de diverse disciplines, dans l’idéal de la construction d’une connaissance intégrée. En effet, le couplage d’approches hétérogènes à différents niveaux et échelles **C :** (Florent) différence ? sera une clé de voute de cette thèse, la moelle épinière de la philosophie sous-jacente et une composante de la théorie qu’on construira.

C : (Florent) non, disent que difficultés existent mais pas lesquelles, et surtout pas dans le champ d’investigation à venir

C : également un développement sur “quanti-quali”

Paradigmes de la complexité en géographie

Pour revenir à notre anecdote introductive, nous nous concentrons sur l’étude d’un objet thématique qui sera les systèmes territoriaux : à l’échelle microscopique, les agents peuvent bien être vus comme éléments constitutifs fondamentaux du territoire, qui émergera comme processus complexe à différentes échelles. Plus généralement, il s’agit par commencer de brosser une revue du rôle de la complexité en géographie. Les géographes sont familiers avec la complexité depuis un certain temps, puisque l’étude des interactions spatiales est l’un de ses objets de prédilection. La variété de champs en géographie (géomorphologie, géographie physique, géographie environnementale, géographie humaine, géographie de la santé, etc. pour en nommer quelques) a sûrement joué un rôle clé dans la constitution d’une

pensée géographique subtile, qui considère des processus hétérogènes et multi-scalaires.

Pumain rappelle dans [Pumain, 2003] une histoire subjective de l'émergence des paradigmes de la complexité en géographie. La cybernétique a produit des théories des systèmes comme celle utilisée par Forrester. **C :** (Florent) pas dvlpé, difficile à lire Plus tard, le glissement vers les concepts de criticalité auto-organisée et d'auto-organisation en physique ont conduit aux développements correspondants en géographie, comme [Sanders, 1992] qui témoigne de l'application des concepts de la synergétique aux dynamiques des systèmes urbains. Enfin, les paradigmes actuels des systèmes complexes ont été introduits par plusieurs entrées. Par exemple, la nature fractale de la forme urbaine a été introduite par [Batty et Longley, 1994] et a eu de nombreuses applications jusqu'à des développements plus récents [Keersmaecker, Frankhauser et Thomas, 2003]. **C :** lier avec les approches de Frankhauser Batty a aussi introduit les automates cellulaires en modélisation urbaine et propose une synthèse jointe avec les modèles basés agents et les fractales dans [Batty, 2007]. **C :** small development on West Bettencourt, scaling and Santa fe school[Bettencourt et al., 2007] Une autre introduction de la complexité en géographie fut pour le cas des systèmes urbains à travers la théorie évolutive des villes de Pumain. En interaction intime avec la modélisation dès ses débuts (le premier modèle Simpop décrit par [Sanders et al., 1997] rentre dans le cadre théorique de [Pumain, 1997]), cette théorie vise à comprendre les systèmes de villes comme des systèmes d'agents adaptatifs en co-évolution, aux interactions multiples, avec différents aspects mis en valeur comme l'importance de la diffusion des innovations. La série des modèles Simpop [Pumain, 2012a] a été conçue pour tester différentes hypothèses de la théorie, comme par exemple le rôle des processus de diffusion de l'innovation dans l'organisation du système urbain. Ainsi, des régimes sous-jacent différents ont été mis en évidence pour les systèmes de ville en Europe et aux Etats-unis [Bretagnolle et Pumain, 2010a]. A d'autres échelles de temps et dans d'autres contextes, le modèle SimpopLocal [Schmitt, 2014] a pour but d'étudier les conditions pour l'émergence de systèmes urbains hiérarchiques à partir d'établissements disparates. Un modèle minimal (au sens de paramètres nécessaires et suffisants) a été isolé grace à l'utilisation de calcul intensif via le logiciel d'exploration de modèles OpenMole [Schmitt et al., 2014], ce qui était un résultat impossible à atteindre de manière analytique pour un tel type de modèle complexe. Les progrès techniques d'OpenMole [Reuillon, Leclaire et Rey-Coyrehourcq, 2013] ont été menés simultanément avec les avances théoriques et empiriques. Les avancées épistémologiques ont également été cruciales dans ce cadre, comme Rey le développe dans [Rey-Coyrehourcq, 2015], et de nouveaux concepts comme la modélisation incrémentale [Cottineau,

Chapron et Reuillon, 2015] ont été découverts, avec de puissantes applications concrètes : [Cottineau, 2014] l'applique sur le système de villes soviétique et isole les processus socio-économiques dominants, par un test systématique des hypothèses thématiques et des fonctions d'implémentation. Des directions pour le développement de de telles pratiques de Modélisation et Simulation en géographie quantitative ont récemment été introduits par Banos dans [Banos, 2013]. Il conclut par neuf principes[4], parmi lesquels on peut citer l'importance de l'exploration intensive des modèles computationnels et l'importance du couplage de modèles hétérogènes, qui sont avec d'autre principes tel la reproductibilité au centre de l'étude des systèmes complexes géographiques selon le point de vue décrit précédemment. Nous nous positionnons dans l'héritage de cette ligne de recherche, travaillant de manière conjointe sur les aspects théoriques, empiriques, épistémologiques et de modélisation.

C : (Florent) point intéressant, mais avant de prendre position pour intégration théorique/empirique, il faut qu'on comprenne pquoi compliqué à fiare (même si hyper riche, déjà des éléments en l'état dans le manuscrit

question de recherche

C : (Florent) logique de dire cela à ce stade mais pas dans manuscrit final La question de recherche et les objets précis sont délibérément flous pour l'instant, puisque nous postulons que la construction d'une problématique ne peut être dissociée de la production d'une théorie correspondante. De manière réciproque, il n'y a aucun sens à poser des questions sorties de nulle part, sur des objets qui ont été seulement partiellement ou brièvement définis. Notre question préliminaire pour entrer dans le sujet, qu'on peut obtenir à partir de cas concrets comme l'anecdote introductive ou la revue de littérature préliminaire, est la suivante :

Comment définir les systèmes territoriaux, et les échelles et ontologies associées, dans une théorie cohérente, innovante et informative sur les processus sous-jacents ? **C :** (Florent) très général et fausse question ! **C :** (Arnaud) Très général, à voir si se tient

Il s'agit bien sûr d'une fausse question à ce stade, mais qui est toujours utile pour diriger la compréhension globale et le lecteur soucieux d'une démarche linéaire classique.

En effet, une caractéristique fondamentale des systèmes territoriaux est leur nature spatio-temporelle, qui est contenue dans leur dynamiques spatio-temporelles. La notion de *processus* au sens de [*Hypergeo*] capture de plus les relations causales entre composantes de ces

4 Je me rappelle René Doursat insister pour la recherche du dernier commandement de Banos

dynamiques, et est ainsi une approche intéressante pour une compréhension voire explication de ces systèmes. L'*échelle* doit être comprise ici au sens opérationnel (caractéristiques physiques) end l'*ontologie* comme les objets réels étudiés[5]. Notre question peut être vue grossièrement comme la recherche de théories et modèles qui révèlent des processus impliqués dans des systèmes complexes contenant aux moins des établissements humains, ce dernier point étant crucial pour la construction d'une problématique convergente plutôt que de se perdre dans des propositions irréalistes et non constructives qui pourrait aller de comprendre tout du cerveau (qui peut être vu comme une brique élémentaire des systèmes territoriaux qui émergent des interactions sociales) à l'écosphère qui inclut aussi les systèmes territoriaux. Ces systèmes spatiaux, que nous préciserons comme *systèmes territoriaux*

C : (Florent) ok bien de préciser cela, mais peut être plus spécifique que de rappeler dimension territoriale (par ex. introduire bifurcations)

CONTENU

This provisory Memoire is organized the following way. A first part with four chapters sets the thematic, theoretical and methodological background. The study of geographical systems implies, because of their complexity, a subtle combination of Theoretical constructions and Empirical Analysis, either in an inductive reasoning or in a didactic constitution of knowledge. The first part aims to approach our subject from the theoretical and methodological point of view, and rather as a *necessary foundation* shall be understood as a body of knowledge *coevolving* with Empirical and Modeling Parts. A linear reading is not necessarily the best way to deeply perceive the implications of theory on empirical and modeling experiments and reciprocally. Some methodological developments are necessary but explicit reference will be done when it will be the case. A first chapter starts from the provisory research question given above and frames from a thematic point of view geographical objects and processes to be studied, resulting in precise research questions. The scene is set up for the construction of our theoretical background in a second chapter, that consists in a geographical theory for territorial systems on the one hand and in an epistemological theory of socio-technical systems **C :** (Florent) c'est quoi ? modeling that frames our approach at a meta-level. **C :**

5 cet usage de la notion d'ontologie biaise naturellement la recherche vers des paradigmes de modélisation puisque qu'elle est proche de celle utilisée dans [Livet et al., 2010a], mais nous prenons la position (développée en détails plus loin) de comprendre toute construction scientifique comme un *modèle*, rendant la frontière entre théories et modèles moins pertinentes que pour des visions plus classiques. Toute théorie doit faire des choix sur les objets décrits, leur relations et les processus impliqués, et contient donc une ontologie dans ce sens.

(Florent) sens ? We then develop methodological considerations on diverse questions implied by theory and required for modeling. Finally, a chapter of quantitative epistemology finishes to pave the way for modeling directions, unveiling literature gaps precisely linked to our question. A second part develops results obtained from empirical analysis and modeling experiments, along with on-going and planned projects in these fields. It first present empirical analysis aimed at identifying stylized facts. Toy-models of urban growth are then proposed, followed by an example and propositions for more complex models. The third part constructs our research objective for the remaining part of our project and sets a corresponding roadmap. Appendices contain non-digest important parts of our work such as models implementation architecture and details and specific tools developed for a reproducible research workflow.

SUR LA LECTURE LINÉAIRE

C : expliquer notre position sur la difficulté d'une présentation linéaire, au delà de faire la synthèse. // bon bouquins y arrivent ? y réfléchir. la métaphore narrative intro/cl parties sera ce squelette linéaire. les deux approches sont compatibles.

Première partie

FOUNDATIONS

This part set up foundations, constructing our research precise subject and questions from a thematic point of view, completed with a theoretical construction for framing at thematic and epistemological levels. We also provide methodological digressions, and a quantitative epistemological analysis completing the manual state of the art. **C :** (Arnaud) ça s'appelle lire

1

INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES

> *Si la question de la priorité de l'œuf sur la poule ou de la poule sur l'œuf vous embarrasse, c'est que vous supposez que les animaux ont été originairement ce qu'ils sont à présent.*
> - Denis Diderot [Diderot, 1965]

Cette analogie est idéale pour introduire les notions de causalité et de processus dans les systèmes territoriaux. En voulant traiter naïvement des questions similaires à notre question de recherche préliminaire, certains on qualifiés les causalités au sein de systèmes complexes comme un problème "de poule et œuf" **C :** (Florent) parler à ce stade de la constroverse Offner 93 **C :** (Arnaud) :) : si un effet semble causer l'autre et réciproquement, comment est-il possible d'isoler les processus correspondants ? Cette vision est souvent présente dans les approches réductionnistes qui ne postulent pas une complexité intrinsèque au sein des systèmes étudiés. L'idée suggérée par Diderot est celle de *co-evolution* qui est un phénomène central dans les dynamiques évolutionnaires des Systèmes Complexes Adaptatifs comme Holland élabore dans [Holland, 2012]. Il fait le lien entre la notion d'émergence (ignorée dans les approches réductionnistes) **C :** (Florent)la encore très epistemo, renforcer connaissance empirique de ces interactions particulieres et en faire état ici , en particulier l'émergence de structures à une plus grand échelle par les interactions entre agents à une échelle donnée, en général concrétisée par un systèmes de limites, qui devient cruciale pour la co-évolution des agents à toutes les échelles : l'émergence d'une structure sera simultanée avec une autre, chacune exploitant leur interrelations et environnements générés conditionnés par le système de limites. Nous explorerons ces idées pour le cas des systèmes territoriaux par la suite. **C :** (Florent)c'est seulement là que tu dis que les syst. territoriaux sont une déclinaison des questionnement précédents

Ce chapitre introductif est destiné à poser le cadre thématique, le contexte géographique sur lesquels les développements suivants se baseront. Il n'est pas supposé être compris comme une revue de littérature exhaustive ni comme les fondations théoriques fondamentales de notre travail (le premier point étant l'objet du chapitre **??** tandis que le second sera traité plus tôt dans le chapitre 9), mais plutôt

comme une construction narrative ayant pour but d'introduire nos objets et positions d'étude, **C :** (Florent) le faire plutot que l'annoncer afin de construire naturellement des questions de recherche précises.

⋆ ⋆

⋆

Ce chapitre est entièrement inédit.

1.1 RÉSEAUX ET TERRITOIRES

1.1.1 *Une circularité naturelle*

TERRITORIALITÉ HUMAINE Une entrée possible dans l'ensemble des objets géographiques que nous proposons d'étudier est la notion de territoire. **C :** (Florent) et un objet de recherche en lui meme En Ecologie, un territoire correspond à l'étendue spatiale occupée par un groupe d'agent ou plus généralement un écosystème. Les *Territoires Humains* sont extrêmement plus complexes de par l'importance de leur représentations sémiotiques, qui jouent un rôle significatifs dans l'émergence des constructions sociétales. **C :** (Florent)pas besoin ni interet de se positionner sur emergence des societes Selon RAFFESTIN dans [RAFFESTIN, 1988], la *Territorialité Humaine* est "la conjonction d'un processus territorial avec un processus informationnel", ce qui implique que l'occupation physique et l'exploitation de l'espace par les sociétés humaines n'est pas dissociable **C :** (Florent) ou est complémentaire ? des représentations (cognitives et matérielles) de ces processus territoriaux, qui influent en retour leur évolution. En d'autres termes, à partir de l'instant où les constructions sociales déterminent la constitution des établissements humains, les structures sociales abstraites et concrètes joueront un role dans l'évolution des systèmes territoriaux, par exemple à travers la propagation d'informations et de représentations, par des processus politiques, ou encore par la correspondance effective entre territoire vécu et territoire perçu. **C :** (Florent)donner exemples concrets serait pédagogique (ex metropole grd paris, cf articles) Bien que cette approche ne donne pas de conditions explicites pour l'émergence d'un système séminal d'établissements agrégés (c'est à dire l'émergence des villes), **C :** (Florent)pourquoi cette interrogation particulière ? elle insiste sur leur role comme lieu de pouvoir et de création de richesse au travers des échanges. Mais la ville n'a pas d'existence sans son hinterland et le système territorial peut difficilement être résumé par ses villes, comme un système de villes. En se restreignant à ce sous-système, il y a toutefois compatibilité entre la théorie de territoires de RAFFESTIN et la théorie évolutive des villes de PUMAIN [PUMAIN, 2010], qui interprète les villes comme des systèmes complexes dynamiques auto-organisés, **C :** (Arnaud) self-organized ? qui agissent comme des médiateurs du changement social : par exemple, les cycles d'innovation s'initialisent au sein des villes et se propagent entre elles. **C :** (Florent)tres pertinent bien sur, mais va aborder la question de l'innovation dans la thèse ? Les villes sont ainsi des agents **C :** (Arnaud) entities ? compétitifs qui co-évoluent (au sens donné précédemment). Le système territorial peut ainsi être compris comme une structure sociale organisée dans l'espace, qui comprend ses artefacts concrets et abstraits. Une étendue spatiale imaginaire

avec des ressources potentielles qui n'aurait jamais connu de contact avec l'humain ne pourra pas être un territoire si elle n'est pas habitée, imaginée, vécue, exploitée, même si ces ressources pourraient être potentiellement exploitée le cas échéant. En effet, ce qui est considéré comme une ressource (naturelle ou artificielle) dépendra de la société (par exemple de ses pratiques et de ses capacité technologiques). [Di Meo, 1998] procède à une analyse historique des différentes conceptions de l'espace (qui aboutissent entre autre à l'espace vécu, l'espace social et l'espace classique de la géographie) et montre comment leur combinaison forme ce que Raffestin décrit comme territoires. Un aspect central des établissements humains qui a une longue tradition d'étude en géographie, et qui est directement relié à la notion de territoire, est celui des *réseaux*. Nous allons voir comment le passage de l'un à l'autre est inévitable et leur définition indissociable. **C :** (Florent)structure generale de l'argumentaire tb, mais devrait expliquer plus en détail ce qu'on appelle réseau (avant de détailler les différents réseaux réel/virtuel, les réseaux ont une inscription spatiale

UNE THÉORIE TERRITORIALE DES RÉSEAUX Nous paraphrasons Dupuy dans [Dupuy, 1987] lorsqu'il propose des éléments pour une "théorie territoriale des réseaux" basée sur le cas concret d'un réseau de transport urbain. Cette théorie présente les *réseaux réels* (i.e. les réseaux concrets , incluant les réseaux de transport) comme la matérialisation de *réseaux virtuels*. **C :** (Florent)dans un second temps seulement, à ce stade "de qui viennent les réseaux" n'est pas une question cruciale, c'est la question réseau/espace/human settlements qui doit être au coeur Plus précisément, un territoire est caractérisé par de fortes discontinuités spatio-temporelles induites par la distribution non-uniforme des agents **C :** (Arnaud) Ontology et des ressources. Ces discontinuités induisent naturellement un réseau de "projets transactionnels" **C :** (Florent)pquoi guillemets ? qui peuvent être compris comme des interactions potentielles entre les éléments du système territorial (agents et/ou ressources). Par exemple, de nos jours les actifs se doivent d'accéder à la ressource qu'est l'emploi, et des échanges économiques s'effectuent entre les différents territoires spécialisés dans les productions de différents types. En tout temps des interactions potentielles ont existé[1] Le réseau d'interaction potentiel est concrétisé quand l'offre s'adapte à la demande, et résulte en la combinaison de contraintes économiques et géographiques avec les motifs de demande, de manière non-linéaire via des agents qu'on peut désigner comme *opérateurs*. Un tel processus est loin d'être immédiat, et conduit à de

1 même quand le nomadisme devait encore être la règle, des réseaux d'interactions potentielles dynamiques dans l'espace ont du exister, mais devaient avoir moins de chance de se matérialiser en des routes matérielles.

forts effets de non-stationarité et de dépendance au chemin **C :** (Florent)Une strategie à adopter serait d'abord de decrire de facon basique, avec exemples concrets, la complexité des interactoins réseau/espace/settelements, puis de rappeler CS et proprietes, puis de decrire lesquelles de ces propriétés presentes dans ces interactions, lequelles modèles vont essayer de reproduire et pquoi. : l'extension d'un réseau existant dépendra de la configuration précédente, et selon les échelles de temps impliquées, la logique et même la nature des opérateurs peut avoir évolué. RAFFESTIN souligne dans sa préface de [OFFNER et PUMAIN, 1996] qu'une théorie géographique articulant espaces, réseaux et territoires n'a jamais été formulée de manière cohérente. **C :** (Florent)redire les ecueils qui sont perçus par Raffestin Il semble que c'est toujours le cas aujourd'hui, même si la théorie évoquée ci-dessus semble être un bon candidat bien qu'elle reste à un niveau conceptuel. La présence d'un territoire humain implique nécessairement la présence de réseaux d'interactions abstraites et de réseaux concrets utilisés pour transporter les individus et les ressources (incluant les réseaux de communication puisque l'information est une ressource essentielle). Selon le régime dans lequel le système considéré se trouve, le rôle respectif du réseau peut être radicalement différent. Selon DURANTON [DURANTON, 1999], les villes pré-industrielles étaient limitées en croissance de par les limitations des réseaux de transport. Les progrès technologiques ont permis de les surmonter **C :** (Florent)trop simplificateur et à mené à la prépondérance du marché foncier dans la formation des villes (et par conséquent un rôle des réseaux de transport qui déterminent les prix par l'accessibilité), et plus récemment à une importance croissante des réseaux de télécommunication ce qui a induit une "tyrannie de la proximité" puisque la présence physique n'est pas remplaçable par une communication virtuelle. Cette approche territoriale des réseaux semble naturelle en géographie, puisque les réseaux sont étudiés conjointement avec des objets géographiques auxquels est associée une théorie, en opposition à la science des réseaux qui étudie brutalement les réseaux spatiaux avec peu de fond thématique [DUCRUET et BEAUGUITTE, 2014]. **C :** (Florent)derniere phrase pas claire **C :** (Arnaud) Ajouter noms ? (biblio ?

DES RÉSEAUX QUI FAÇONNENT LES TERRITOIRES ? Cependant les réseaux ne sont pas seulement une manifestation matérielle de processus territoriaux, mais jouent également leur rôle dans ces processus comme leur évolution peut influencer l'évolution des territoires en retour. Dans le cas des *réseaux techniques*, une autre désignation des réseaux réels donnée dans [OFFNER et PUMAIN, 1996], de nombreux exemples de tels retroactions peuvent être mis en évidence : l'interconnexion des réseaux de transport permet des motifs de mobilité multi-échelles, **C :** (Florent)chose plus basiques à dire

en premier (favorise croissance urbaine) formant ainsi le territoire vécu. A une plus petite échelle, des changements de l'accessibilité peuvent induire l'adaptation d'un espace fonctionnel urbain. Il emerge alors une difficulté intrinsèque : **C :** (Florent) TB mais en parler avant, c'est cela le coeur il est loin d'évident d'attribuer des mutations territoriales à une évolution du réseau and réciproquement la matérialisation d'un réseau à des dynamiques territoriales précises. Revenir à la citation de Diderot devrait aider à ce point, au sens où il ne faut pas considérer le réseau ni les territoires comme des systèmes indépendants qui s'influenceraient mutuellement par des relations causales, mais comme des composantes fortement couplées d'un système plus large. La confusion autour de possibles relations causales simples a nourri un débat scientifique encore actif aujourd'hui. Les méthodologies pour identifier ce qui est nommé *effets structurants* des réseaux de transport ont été proposées par les planificateurs dans les années 1970 [Bonnafous et Plassard, 1974; Bonnafous, Plassard et Soum, 1974]. Il aura fallu un certain temps pour un positionnement critique sur l'usage non raisonné et decontextualisé de ces méthodes par les planificateurs et les politiques qui les mobilisaient généralement pour justifier des projets de transports de manière technocratique. Cela a été fait en premier par Offner dans [Offner, 1993]. Récemment un édition spéciale du même journal sur ce débat [L'Espace géographique, 2014] a rappelé d'une part que les mauvaises interprétations et les mauvais usages étaient encore largement présent aujourd'hui dans les milieux opérationnels de la planification comme [Crozet et Dumont, 2011a] confirme, et d'autre part qu'il faudrait encore une certaine quantité de progrès scientifique pour comprendre en profondeur les relations entre réseaux et territoires. Les débats récents en juillet 2017 relatifs à l'ouverture des LGV Bretagne et Sud-Ouest ont montré toute l'ambiguïté des positions, des conceptions, des imaginaires à la fois des politiques mais aussi du public : refus du financement d'élus qui s'attendaient au prolongement vers Toulouse et l'Espagne, spéculation dans les quartiers de gare, questionnements des pratiques de mobilité quotidienne mais aussi sociale. La complexité et la portée des sujets montre bien la difficulté d'une compréhension systématique d'effets du transport sur les territoires. Pumain souligne que des travaux récents ont révélé des effets systématiques sur de très longues échelles temporelles (comme e.g. le travail de Bretagnolle sur l'évolution des chemins de fer, qui montre une sorte d'effet structurel sur la nécessité de connexion au réseau des villes, afin de rester actives, mais qui n'est ni suffisant ni totalement causal). **C :** (Florent)développer ce genre de categories macro c'est très interessant A un niveau macroscopique des motifs typiques d'interaction émergent, mais les trajectoires microscopiques du systèmes sont essentiellement chaotiques : la compréhension des dynamiques couplées dépend fortement de l'échelle considérée. A

une petite échelle il est peu raisonnable de vouloir montrer des comportement systématiques, comme le rappelle OFFNER. Par exemple, sur des territoires de montagne français comparables, [BERNE, 2008] montre que les réactions à un même contexte d'évolution du réseau de transport peut mener à des réactions territoriales très diverses, certains trouvant de forts bénéfices par la nouvelle connectivité, d'autres au contraire devenant plus fermés. Ces retroactions potentielles des réseaux sur les territoires n'agit pas nécessairement sur des composantes concretes : CLAVAL montre dans [CLAVAL, 1987] que les réseaux de transport et de communication contribuent à la représentation collective d'un territoire en agissant sur un sentiment d'appartenance. **C :** (Florent) la encore de second ordre, a ressortir pour lutetia

SYSTÈMES TERRITORIAUX Ce voyage des territoires aux réseaux, et retour, nous permet d'esquisser une définition préliminaire d'un système territorial sur laquelle se basera les considérations théoriques suivantes. **C :** (Florent)si c'est autant au coeur, présenter avant Comme nous avons mis en exergue le rôle des réseaux, la définition se doit de les prendre en compte.

Définition provisoire. *Un Système Territorial est un territoire humain auquel peuvent être associés à la fois un réseau d'interactions et un réseau réel. Les réseaux réels sont une composante à part entière du système, jouant dans les processus d'évolution, au travers de multiples retroactions avec les autres composantes à plusieurs échelles spatiales et temporelles.*

C : (Florent) feedback : propriété, pas def ; plus une axiomatique qu'une demo ?

Cette lecture des systèmes territoriaux est conditionnée à l'existence des réseaux et pourrait écarter certains territoires humains, mais il s'agit d'un choix délibéré justifié par les considérations précédentes, et qui précise notre sujet vers l'étude des interactions entre réseaux et territoires. **C :** (Florent) formulé comme ça, on peut penser que network pas inclus dans territoire

reflexion en parcourant toutes les approches luti : le transport est différent des réseaux de transports, il correspond à l'utilisation de celui-ci par les agents territoriaux- donc ceux ci ne font pas partie du territoires à proprement parler. transport demand, offer (very limited), congestion etc. : more at the scale of mobility, pas vraiment notre concern du coup.

1.1.2 *Réseaux de Transport*

LA PARTICULARITÉ DES RÉSEAUX DE TRANSPORT Déjà évoqués dans le cas des effets structurants des réseaux, les réseaux de transports jouent un rôle central dans l'évolution des territoires, mais il n'est évidemment pas question de leur attribuer des effets causaux

déterministes. Même si d'autres types de réseaux sont également fortement impliqués dans l'évolution des systèmes territoriaux (voir e.g. les débats sur l'impact des réseaux de communication sur la localisation des activités économiques), les réseaux de transport conditionnent d'autres types de réseaux (logistique, échanges commerciaux, interactions sociales concrètes pour donner quelques exemples) and semblent dominer dans les motifs d'évolution territoriale, en particulier dans nos sociétés contemporaines qui sont devenues dépendantes des réseaux de transport [Bavoux et al., 2005]. Le développement du réseau français à grande vitesse est une illustration pertinente de l'impact des réseaux de transport sur les politiques de développement territorial. Présenté comme une nouvelle ère de transport sur rail, une planification par le haut de lignes totalement nouvelles et indépendantes de par leur vitesse deux fois plus élevée, a été présenté comme central pour le développement [Zembri, 1997]. Le manque d'intégration de ces nouveaux réseaux avec l'existant et avec les territoires locaux est à présent observé comme une faiblesse structurelle et des impacts négatifs sur certains territoires ont été prouvés [Zembri, 2008]. Une revue faite dans [Bazin et al., 2011] confirme qu'aucune conclusion générale sur des effets locaux d'une connection à une ligne à grande vitesse ne peut être tirée, bien que ce sésame garde une place conséquente dans les imaginaires des élus. Ces exemples illustrent comment les réseaux de transport peuvent avoir des effets à la fois directs et indirects sur les dynamiques territoriales. Le développement des différentes Lignes à Grande Vitesse s'inscrit dans des contextes territoriaux très différents, et il est dans tous les cas délicat de penser pouvoir interpréter des processus hors de ceux-ci : par exemple, les lignes LGV Nord et LGV Est s'inscrivent dans des échelles européennes plus vastes que la LGV Bretagne ouverte en juillet 2017. La planification intégrée, au sens d'une planification coordonnée entre les infrastructures de transport et le développement urbain, considère le réseau comme une composante déterminante du système territorial. Les Villes Nouvelles parisiennes sont un tel cas qui témoigne de la complexité de ces actions de planification qui le plus souvent ne mène pas au effets initialement désirés [Ostrowetsky, 2004]. Des projets récents comme [L'Hostis, Soulas et Wulfhorst, 2012] ont tenté d'implémenter des idées similaires, mais il manque pour l'instant de recul pour juger de leur succès à produire un territoire effectivement intégré. **C :** (Florent) dans le detail, quels sont les ordres de grandeur des temps pour que les réseaux puissent avoir un effet ? Les réseaux de transports sont dans tous les cas au centre de ces approches des territoires urbains. Nous nous concentrerons par la suite sur les réseaux de transport **C :** (Florent)tous ? pour toutes ces raisons évoquées ici.

TRANSPORTS ET ACCESSIBILITÉ [MILLER, 1999] on three different way to approach accessibility : time-geography and constraints, user utility based measures, and transportation time. It derives measures for each in perspective of WEIBULL's axiomatic frameworks and reconcile the three in a way.

La notion d'accessibilité surgit rapidement lorsqu'on s'intéresse aux réseaux de transport. Basée sur la possibilité d'accéder un lieu par un réseau de transport (pouvant prendre en compte la vitesse, la difficulté de se déplacer), elle est généralement définie comme un potentiel d'interaction spatiale[2] [BAVOUX et al., 2005]. Cet objet est souvent utilisé comme un outil de planification ou comme une variable explicative de localisation des agents par exemple. **C :** (Florent) dire d'abrd à quoi peut servir Il faut cependant rester prudent sur son usage inconditionnel. Plus précisément, il peut s'agir d'une construction qui ignore une partie conséquente des dynamiques territoriales. La mystification **C :** (Florent) trop fort, Hadri montre que étude et prod de l'infra sont pas indep, mais pas de myst **C :** (Arnaud) Contexte français de la notion de *mobilité* a été montrée par COMMENGES dans [COMMENGES, 2013b], qui révèle que la majorité des débats sur la modélisation de la mobilité et les notions correspondantes était majoritairement construites de manière ad-hoc par les administrateurs de transports issus du *Corps des Ponts* **C :** (Florent) lecture trop rapide qui importaient brutalement les outils et méthodes des Etats-Unis sans adaptation ni reflexion adaptée au contexte français. L'accessibilité pourrait de même être une construction sociale et n'avoir que peu de fondement théorique, puisqu'il s'agit en grande partie d'un outil de modélisation et de planning. Les débats récents sur la planification du *Grand Paris Express* [MANGIN, 2014], **C :** (Florent) interessant : à creuser cette nouvelle infrastructure de transport métropolitaine planifiée pour les vingts prochaines années, a révélé l'opposition entre une vision de l'accessibilité comme un droit pour les territoires désavantagés, contre l'accessibilité comme un moteur du développement économique pour des zones déjà dynamiques, les deux étant difficilement compatibles car correspondent à des couloirs de transport très différents. De tels problèmes opérationnels confirment la complexité du rôle des réseaux de transports dans les dynamiques des systèmes territoriaux, et nous devrons donner dans notre travail des éléments de réponse pour une définition de l'accessibilité qui intégrerait les dynamiques territoriales intrinsèques.

ECHELLES ET HIERARCHIES Un aspect incontournable des réseaux de transport que nous devrons prendre en compte dans nos dévelop-

2 et souvent généralisée comme une *accessibilité fonctionnelle*, par exemple les emplois accessibles aux actifs d'un lieu. Les potentiels d'interaction spatiaux s'exprimant dans les lois de gravité peuvent aussi être compris de cette façon.

pements futurs et la hiérarchie. Les réseaux de transport sont par essence hiérarchique, dépendant des échelles dans lesquelles ils sont intégrés. [Louf, Roth et Barthelemy, 2014] montre empiriquement des propriétés de loi d'échelle pour un nombre conséquent d'aires métropolitaines à travers la planète, et les lois d'échelle révèlent la présence de hiérarchie dans un système, comme pour la hiérarchie de taille dans les systèmes de villes exprimée par la loi de Zipf [Nitsch, 2005] ou d'autres lois d'échelle urbaines [Arcaute et al., 2013 ; Bettencourt et Lobo, 2015]. La topologie du réseau de transport a été montrée suivre de telles lois pour la distribution de ses mesures locales comme la centralité [Samaniego et Moses, 2008]. **C :** (Florent) tb mais comment relie à partie juste avant ? La hiérarchie semble jouer un rôle particulier dans les processus d'interaction, comme Bretagnolle [Bretagnolle, 2009a] a souligné une correlation croissante dans le temps entre la hiérarchie urbaine et la hiérarchie de l'accessibilité temporelle pour le réseau ferroviaire français, **C :** (Florent) tb mais séparé entre réseau et pop ; pourquoi pas regarder la hiérarchie de l'access ? marqueur de retroactions positives entre le rang urbain et la centralité de réseau. Différents régimes dans le temps et l'espace ont été identifiés : pour l'évolution du réseau ferroviaire français e.g., une première phase d'adaptation du réseau à la configuration urbaine existante a été suivie par une phase de co-évolution i.e. au sens où les relations causales sont devenues difficiles à identifier. L'impact de la contraction de l'espace-temps par les réseaux sur le potentiel de croissance des villes avait déjà été montré pour l'Europe par des analyses exploratoires dans [Bretagnolle, Pumain et Rozenblat, 1998]. L'evolution du réseau ferroviaire aux Etats-unis a suivi une dynamique bien différente, sans diffusion hiérarchique, donnant forme localement à la croissance urbaine. **C :** (Florent) un peu rapide mais dans l'autre sens : cela n'a pas marché partout mais contexte particulier de la conquete de l'ouest est intéressant à souligner Cela met l'emphase sur la présence de dépendance au chemin **C :** (Florent) en parler avant si c'est le coeur du projet pour les trajectoires des systèmes urbains : la présence en France d'un système préalable de villes et de réseau (routes postales) a fortement influencé le développement du réseau ferré, tandis que son absence aux Etats-unis a conduit à une histoire complètement différente. Une question ouverte est si des processus génériques sont implicites aux deux évolutions, chacun correspondant à des réalisations différentes avec des conditions initiales et des méta-paramètres différentes (des *régimes* différents au sens des transitions des systèmes de peuplement introduites dans le projet de recherche courant ANR TransMonDyn, puisque une transition peut être comprise comme un changement de stationnarité des méta-paramètres **C :** (Florent) trop rapide ce n'est pas compréhensible en l'état d'une dynamique générale). En termes de systèmes dynamiques, cela revient à se demander si les dy-

namiques des ensembles de catastrophe (composantes à plus grandes échelles temporelles) obéissent à des équations similaires que la position et nature des attracteurs pour un système dynamique stochastique qui donnent son régime courant, en particulier si le système est dans un état local divergent (exposant de Liapounov local positif) ou en train de converger vers des mécanismes stables [Sanders, 1992]. Pour répondre à cette question en même temps que l'isolation des processus de co-évolution pour ce régime, [Bretagnolle, 2009a] propose la modélisation comme élément de réponse constructif. Nous verrons dans le chapitre suivante comme la modélisation peut être source de connaissance sur les processus territoriaux.

sur le TOD : example multi-critère [L'Hostis, SOULAS et Vulturescu, 2016] → ajouter paragraphe planification ?

le tod : litérature planing, design etc, coordination tranposrt urba ; citer mais après chap 2 plus notre concern (comprendre et pas planifier ; fait alors partie des processus abstraits)

exemple tod à Lille [Liu et Alain, 2014], le cas Français.

TRANSPORTS ET MOBILITÉ sur la mobilité : nos questionnements à une autre échelle ? cf [Fusco, 2004] relations causales

1.1.3 *Interactions entre Réseaux et Territoires*

At this state of progress, we have naturally identified a research subject that seems to take a significant place in the complexity of territorial systems, that is the study of interactions between transportation networks and territories. In the frame of our preliminary definition of a territorial system, this question can be reformulated as the study of networked territorial systems with an emphasize on the role of transportation networks in system evolution processes.

C : (Florent) ok : à quelles échelles de temps et d'espace se place t'on (même un intervalle)

- ici donner des exemples concrets -

Gaelle Lesteven Metro toulouse

[Fritsch, 2007] tramway de Nantes et densification

C : (Florent) aéroport MCR : Ciudad Real

[Bazin, Beckerich et Delaplace, 2007] prix immobilier à Reims : effet du TGV très localisé seulement, dynamique globale. [Bazin et al., 2010] conclusion idem revue littérature incluant grise.

impact sur le long terme : systematic review of empirical studies [Kasraian et al., 2016] ; [Rietveld, 1994] similar older from economics viewpoint

CO-ÉVOLUTION DES RÉSEAUX ET DES TERRITOIRES On se place déjà dans l'idée d'une co-evolution - introduire le concept selon exemples empiriques et littérature.

[Offner, 1993] parle de congruence - à lier avec vision systémique de l'époque - serait une vision préliminaire de la co-évolution.

1.2 DE PARIS À ZHUHAI

1.2.1 *Le Grand Paris : histoire et enjeux*

La région parisienne est une bonne illustration de la complexité des interactions entre réseaux de transports et territoires, au cours du temps et à l'échelle intermédiaire d'une région métropolitaine globalement mono-centrique.

[GILLI et OFFNER, 2009] propose en 2009 un diagnostic de la situation institutionelle de la région parisienne, et des pistes pour une approche couplée entre gouvernance et aménagement. La préfiguration de "l'instauration d'un acteur collectif métropolitain" correspond à la métropole du Grand Paris qui sera inaugurée 7 ans plus tard

[PADEIRO, 2012]

1.2.2 *Le Delta de la Rivière des Perles : nouveaux régimes urbains et Mega-City Regions*

TODO : *some "comparable" maps would be useful : ask Chenyi most precise data on PRD : territorial variables and transportation networks ?*

Parler du pont et des bifurcations induites (cf intro chap 5)

Si la notion de megalopolis peut être tracée jusqu'à GOTTMANN [GOTTMANN, 1964], et qu'elle est à l'origine de celle de Mega-city Region consacrée par HALL [HALL et PAIN, 2006], il est clair que cette dernière est toujours plus d'actualité avec l'apparition récente de nouveaux régimes, notamment par l'urbanisation croissante dans des pays à forte croissance et en mutation très rapide comme la Chine [SWERTS et DENIS, 2015].

1.2.3 *Comparabilité des études de cas*

1.3 ELEMENTS DE TERRAIN

1.3.1 *Une Experience en Observation Flottante*

Si le diable est dans les détails, les systèmes de transport entre autres sont l'allégorie de cette adage. Ce que certains appellent détail contient la majorité de l'information pour d'autres. Logiquement enfermés dans une bulle scientifique, malgré toutes les volontés développées en introduction, on tâchera de rester conscient de la nature et la portée de la connaissance produite ici. Ce que nous pourrions appeler détail, lors de l'étude de l'accessibilité d'un réseau de transport par exemple, tel des impressions ressenties par les usagers ou les relations sociales induites par les situations découlant des dynamiques du systèmes, seront le centre du questionnement pour un anthropologue ou sociologue. Une telle connaissance, qui trouverait certainement une place dans nos problématiques, est hors de notre portée de par l'absence de *terrain* de longue durée. Nous proposons toutefois ici d'ébaucher une entrée qualitative d'un certain type, pour suggérer une façon de compléter nos connaissances.

L'entrée prise suit la méthode *d'observation flottante*, introduite à l'interface de l'anthropologie et la sociologie par [PÉTONNET, 1982], avec l'ambition de fonder une anthropologie urbaine. Il ne s'agit pas exactement de la même idée que l'anthropologie de l'espace de *Choay* Répondant à un besoin de mouvement que le sédentaire éprouve facilement, le chercheur se place au centre du processus de production de connaissances, nous citons, en "rest[ant] en toute circonstance vacant et disponible, à ne pas mobiliser l'attention sur un objet précis, mais à la laisser flotter afin que les informations la pénètrent sans filtre, sans a priori, jusqu'à ce que des points de repère, des convergences, apparaissent et que l'on parvienne alors à découvrir des règles sous-jacentes". Sans s'y méprendre et considérer la méthode comme une négligence méthodologique, nous y voyons une opportunité d'un accès rapide et à faible coût dans le monde du qualitatif, tout en restant conscient de sa portée très limitée. La disposition d'esprit peut être rapprochée de la philosophie La méthode peut servir d'étude préliminaire pour fixer des protocoles et grilles précises d'entretien : elle est par exemple utilisée justement au sujet du transport par [ALBA et AGUILAR, 2012].

Les mouvements pendulaires à échelle moyenne sont nécessairement vécus d'une façon particulière en comparaison à d'autres lieux géographiques et à d'autres échelles sur le même lieu. Et si une façon d'appréhender des faits stylisés particuliers était alors d'effectuer l'analogue d'une étude de perturbation sur le système, mais en prenant comme référentiel l'observateur lui-même ? Il s'agirait de faire porter un choc sur une situation "d'équilibre", puis de se laisser flotter au gré du courant pour appréhender la réaction et certains mé-

canismes qu'il aurait été difficile de considérer en suivant sa routine. Une expérience naturelle causée par une perturbation des transports (qui en région francilienne est bien courante) est un événement déclencheur de "naufrages" de l'observation, au sens où le chercheur peut capturer des situations et réactions individuelles particulières.

AU-DELÀ DU CHARLATANISME : SYSTÉMATISER LA MÉTHODE FLOTTANTE Notre méthodologie est relativement simple : se laisser errer dans les transports en commun, avec ou sans but et de manière ou non aléatoire, mais en essayant sur chaque trajet de maximiser les opportunités de mise en situation ou de capture d'évènement. La répétition de l'expérience visera également à maximiser la couverture spatiale, temporelle, de situation. Une production traçable est nécessaire à chaque itération, qu'il s'agisse de description factuelle, de description perçue, de semi-synthèse

1.3.2 *Entretiens*

1.3.3 *Analyse Urbanistique*

Le ciel est gris et les visages fermés, Oxmo avait tristement raison, ce Soleil du Nord n'avait de lumière que le nom. L'initié ne saura s'y tromper et ressentira au fond de lui-même cette banale routine d'un aller-retour quotidien en RER. Il ne cherchera ni à maudire les planifications successives dont les stratifications temporelles ont laissé décanter cette organisation territoriale incongrue, ni à se prendre à rêver d'une trajectoire de vie alternative puisque choisir c'est un peu mourir et qu'il ne se sent pas une âme de Phoenix aujourd'hui. Peut être que la beauté de la ville est finalement dans ces tensions qui la façonnent à tous les niveaux et dans tous les domaines, ces paradoxes qui deviennent cadre de vie au point d'asséner quotidiennement une vérité. Cette philosophie de couloir de métro, le francilien en fait son cheval de bataille car après tout s'il vit en ville il doit bien la connaître. Encore un rail cassé sur le A, "tout cela est mal géré, et ce réseau est mal conçu" vocifère un utilisateur journalier, s'improvisant expert en planification ; d'autres plus patients prennent leur mal en patience mais se présentent tout aussi connaisseurs d'une illusoire vision d'ensemble d'un territoire aux multiples visages. Ces usagers *sont* pourtant le système, de manière concrète à leur échelle d'espace et de temps, par induction et émergence aux échelles supérieures. La fourmi est supposée ne pas avoir conscience de l'intelligence collective dont elle est une des composantes fondamentales. Ils n'ont de la même manière que peu de perception de l'auto-désorganisation dont ils sont la source, peut-être la cause, et qui très sûrement subissent les désagréments de ses dynamiques. Se laisser flotter dans les transports franciliens est une expérience intemporelle. Presque thérapeutique parfois, quand l'un commence à perdre son optimisme quant à l'intérêt d'une vie urbaine, une excursion aléatoire en métro rappelle rapidement la richesse et la diversité qui sont un des plus grand succès des villes. C'est cette variété apparente de profils que le chercheur retiendra principalement de ces errements dont la méthodologie est de ne pas avoir de méthodologie, et il gardera à l'esprit qu'il n'existe pas d'échelle où un traitement spécifique de chaque objet géographiques n'est pas nécessaire : en quelque stations sur la ligne 4 le profil des quartiers et donc des usagers change profondément et souvent sans transition au moins trois fois, comme sur la ligne 13 nord où les motifs horaires soulignent d'autant plus de dures réalités socio-économiques qui sont en fait géographiques dans cet *espace produit* de la métropole. Lorsqu'il s'agit de modéliser, prendre en compte les limites de toute tentative de généralisation est d'autant plus cruciale comme chaque modèle est un équilibre fragile entre spécificité et généralité. ENCADRÉ : *Une expérience en observation flottante en région parisienne*

ENCADRÉ : *Une expérience en observation flottante, Guangdong, Zhuhai*

CONCLUSION DU CHAPITRE

2

MODÉLISER LES INTERACTIONS ENTRE RÉSEAUX ET TERRITOIRES

Si la littérature empirique et thématique, ainsi que les cas d'études développés précédemment, semblent converger vers un consensus sur la complexité des relations entre réseaux et territoires, et dans certaines configurations et à certaines échelles de relations circulaires causales entre dynamiques territoriales et dynamiques des réseaux de transports que l'on se proposera de désigner par *co-évolution*, ceux-ci semblent diverger sur toute explication potentiellement simple ou systématique, comme le rappelle par exemple les débats autour des effets structurants des infrastructures [Offner, 1993]. Au contraire, les multiples situations géographiques poussent à privilégier des études ciblées très fortement dépendantes du contexte et du travail de terrain. Or l'explication géographique et la compréhension des processus est très vite limitée dans cette approche, et intervient un besoin d'un certain niveau d'abstraction et de généralisation. C'est sur un tel point que la Théorie Evolutive des Villes est absolument remarquable, puisqu'elle arrive à combiner des schémas et modèles généraux aux particularités géographiques, et en tire même parti, tandis que certaines théories issues de la physique comme la Théorie du Scaling de West [West, 2017] peuvent être plus difficile à digérer pour les géographes de par leur positionnement d'universalité qui est à l'opposé de leurs épistémologies habituelles. Dans tous les cas, le *medium* qui permet de gagner en généralité sur les processus et structures des systèmes est toujours le *modèle* (voir 9.3 pour un développement des domaines de connaissance et du rôle du modèle). Comme le rappelle J.P. Marchand [Raimbault, 2017d], "notre génération a compris qu'il y avait une co-évolution, la votre cherche à la comprendre", ce qui appuie le pouvoir de compréhension apporté par la modélisation et la simulation qui pourraient être aujourd'hui à leur balbutiements. Sans développer les innombrables fonctions que peut avoir un modèle, nous nous baserons sur l'adage de Banos qui soutient que "modéliser c'est apprendre", et suivant notre positionnement dans une science des systèmes complexes suggéré en introduction, nous ferons ainsi de la *modélisation des interactions entre réseaux et territoires* notre principal sujet d'étude, outil, objet (même si dans une lecture rigoureuse de 9.3 ce positionnement n'a pas de sens puisque notre démarche contenait déjà des modèles à partir du moment où elle était scientifique). Ce chapitre peut être vu comme un "état de l'art" des démarches de modélisation des interactions entre réseaux et territoires, mais vise à être aussi objectif et exhaustif que

possible : pour cela, nous mobiliserons des analyses en épistémologie quantitative. Dans une première section 2.1, nous revoyons de manière interdisciplinaire les modèles pouvant être concernés, même de loin, sans a priori d'échelle temporelle ou spatiale, d'ontologies, de structure, ou de contexte d'application. Les modèles de changement d'usage du sol très appliqués en planification sont tout autant concernés que des modèle totalement abstraits issus de la biologie ou de la physique, que des approches intégrées en géographie ou spécifiques en économie. Cet aperçu suggère des structures de connaissances assez indépendantes et des disciplines ne communiquant que rarement. Nous procédons à une revue systématique algorithmique dans 2.2 pour reconstruire leur paysage scientifique, dont les résultats tendent à confirmer ce cloisonnement. L'étude est complétée par une analyse d'hyperréseau, combinant réseau de citation et réseau sémantique issu d'analyse textuelle, qui permet de mieux cerner les relations entre disciplines, leur champs lexicaux et leur motifs d'interdisciplinarité. Cette étude permet la constitution du corpus utilisé pour la modélographie et la meta-analyse effectuée en dernière section 2.3, qui dissèque la nature d'un certain nombre de modèles et la relie au contexte disciplinaire, ce qui pose les bases et le cadre précis des efforts de modélisation qui seront développés par la suite.

⋆ ⋆

⋆

*Ce chapitre est inédit pour sa première section ; reprend dans sa deuxième section le texte traduit de [*Raimbault, *2015a], puis pour sa deuxième partie la méthodologie de [*Raimbault, *2016c], les outils de [*Bergeaud, Potiron *et* Raimbault, *2017a] et des passages de [] ; et est enfin inédit pour sa dernière partie.*

2.1 MODÉLISER LES INTERACTIONS

2.1.1 *Modélisation en Géographie Quantitative*

La modélisation joue en Géographie Théorique et Quantitative (TQG) un rôle fondamental. CUYALA procède dans [CUYALA, 2014] à une analyse spatio-temporelle du mouvement de la Géographie Théorique et Quantitative en langue française et souligne l'émergence de la discipline comme une combinaison d'analyses quantitatives (e.g. analyse spatiale et pratiques de modélisation et de simulation) et de construction théoriques. On peut dater à la fin des années 70 cette dynamique, profondément liée à l'utilisation et l'appropriation des outils mathématiques [PUMAIN et ROBIC, 2002]. L'intégration de ces deux composantes permet la construction de théories à partir de faits stylisés empiriques, qui produisent à leur tour des hypothèses théoriques pouvant être testées sur les données empiriques. Cette approche est née sous l'influence de la *New Geography* dans les pays Anglo-saxons et en Suède. Une histoire étendue de la genèse des modèles de simulation en géographie est faite par REY dans [REY-COYREHOURCQ, 2015] avec une attention particulière pour la notion de validation de modèles. L'utilisation de ressources de calcul pour la simulation de modèles est antérieur à l'introduction des paradigmes de la complexité, remontant par exemple à FORRESTER, informaticien qui a été pionnier des modèles d'économie spatiale inspirés par la cybernétique. Avec l'augmentation des potentialités de calcul, des transformations épistémologiques ont également suivi, avec l'apparition de models explicatifs comme outils expérimentaux. REY compare le dynamisme des années soixante-dix quand les centres de calcul furent ouverts aux géographes à la démocratisation actuelle du Calcul Haute Performance (calcul sur grille à l'utilisation transparente, voir [SCHMITT et al., 2014] pour un exemple des possibilités offertes en terme de calibration et de validation de modèle, réduisant le temps de calcul nécessaire de 30 ans à une semaine - ces techniques jouent un rôle clé pour les résultats que nous obtiendrons par la suite), qui est également accompagnée par une évolution des pratiques [BANOS, 2013] et techniques [CHÉREL, COTTINEAU et REUILLON, 2015] de modélisation. La modélisation, et en particulier les modèles de simulation, est vue par beaucoup comme une brique fondamentale de la connaissance : [LIVET et al., 2010a] rappelle la combinaison des domaines empirique, conceptuel (théorique) et de la modélisation, avec des retroactions constructives entre chaque. Une modèle peut être un outil d'exploration pour tester des hypothèses, un outil empirique pour valider une théorie sur des jeux de données, un outil explicatif pour révéler des causalités et ainsi des processus internes au système, un outil constructif pour construire itérativement une théorie conjointement avec celle des modèles associés. Ce sont des exemples de fonctions

parmi d'autres : Varenne donne dans [Varenne, 2010b] une classification raffinée des diverses fonctions d'un modèle. Nous considérons la modélisation comme un instrument fondamental de connaissance des processus au sein de systèmes complexes adaptatifs, et précisons encore notre question de recherche, qui s'intéressera aux *modèles impliquant des interactions réseaux et territoires*.

MODÉLISATION ET EQUILIBRE Lorsqu'on se détache des approches proposées par l'Economie géographique, la plupart des approches en Géographie Théoriques et Quantitatives sont généralement basées sur des hypothèses de systèmes hors-équilibre [Pumain, 2017]. Les premières contributions de la théorie de Prigogine à l'étude des systèmes urbains, comme par exemple les modèles d'entropie de Allen comme celui étudié par [Pumain, Saint-Julien et Sanders, 1984], ont permis de consacrer les ontologies de l'auto-organisation dans des modèles formels, puis plus tard de simulation, pour les dynamiques urbaines. Les modèles que nous considérerons par la suite rentreront a priori dans cette catégorie pour leur grande majorité. Si un équilibre est supposé à certaines échelles d'espace ou de temps, ce sera souvent dans un contexte de déséquilibre au niveau supérieur, et donc de non-stationnarité du premier niveau.

2.1.2 *Modéliser les territoires et réseaux*

Au sujet de notre question précise des interactions entre réseaux de transport et territoires, nous proposons un aperçu des différentes approches. Selon [Bretagnolle, Paulus et Pumain, 2002], "*les idées des spécialistes de la planification cherchant à donner des définitions des systèmes de ville, depuis 1830, sont étroitement liées aux transformations des réseaux de communication*". C'est en quelque sorte la prophétie autoréalisatrice inversée, au sens où elle est déjà réalisée avant d'être formulée. Cela implique que les ontologies et les modèles correspondants proposés par les géographes et les planificateurs sont fortement liés aux préoccupations historiques courantes, ainsi forcément limités en portée et raisons. Au delà de la question de la définition du système sur laquelle nous reviendrons maintes fois, on comprend bien l'impact que peut avoir cette influence sur la portée des modèles développés. Dans une vision perspectiviste de la science [Giere, 2010c] de telles limites sont l'essence de l'entreprise scientifique, et comme nous démontrerons en chapitre 9 leur combinaison et couplage dans le cas de modèles est une source de connaissance.

Modèles LUTI

Un partie importante de la littérature proposant des modélisations des interactions entre réseaux et territoires se trouve dans le domaine de la planification urbaine, avec les *modèles d'interaction entre usage du*

sol et transport (*LUTI*). Ces travaux peuvent être difficiles à cerner car liés à différentes disciplines. Par exemple, du point de vue de l'Economie Urbaine, les propositions de modèle intégrés existent depuis un certain temps [Putman, 1975]. La variété des modèles existants a conduit à des comparaisons opérationnelles [Paulley et Webster, 1991]. Plus récemment, les avantages respectifs des approches statiques et dynamiques a été étudié par [Kryvobokov et al., 2013], dans un cadre métropolitain sur des échelles de temps moyennes. Dans tous les cas, ce type de modèle opère généralement à des échelles temporelles et spatiales relativement faibles. [Wegener et Fürst, 2004] donne un état de l'art des études empiriques et de modélisation sur ce type d'approche des interactions entre usage du sol et transport. Le positionnement théorique est plutôt proche des disciplines de la socio-économie des transports et de la planification (voir les paysages dressés en 2.2), et pas forcément proche de nos raisonnements géographiques qui se veulent de comprendre également des processus sur le temps long. Pas moins de dix-sept modèles sont comparés et classifiés, parmi lesquels aucun n'inclut une évolution endogène du réseau de transport sur les échelles de temps relativement petites des simulations. Une revue complémentaire est faite par [Chang, 2006], élargissant le contexte avec l'inclusion de classes plus générales de modèles, comme des modèles d'interactions spatiales (parmi lesquels l'attribution du traffic et les modèles à quatre temps), les modèles de planification basés sur la recherche opérationnelle (optimisation des localisations), les modèles microscopiques d'utilité aléatoire, et les modèles de marché foncier. Différents aspects du même système peuvent être traduits par divers modèles (comme e.g. [Wegener, Mackett et Simmonds, 1991]), et le traffic, les dynamiques résidentielles et d'emploi, l'évolution de l'usage du sol en découlant, influencée aussi par un réseau de transport statique, sont généralement pris en compte. Toutes ces techniques opèrent également à une petite échelle et considèrent au plus l'évolution de l'usage du sol. [Iacono, Levinson et El-Geneidy, 2008] couvre un horizon similaire avec une emphase supplémentaire sur les modèles à automates cellulaires d'évolution d'usage du sol et les modèles basés agent. Les modèles LUTI sont toujours largement étudiés et appliqués, comme par exemple [Delons, Coulombel et Leurent, 2008] qui est utilisé pour la région métropolitaine parisienne. La courte portée temporelle d'application de ces modèles et leur nature opérationnelle les rend utiles pour la planification, ce qui est assez loin de notre souci d'obtenir des modèles explicatifs de processus géographiques. En effet, il est souvent plus pertinent pour un modèle utilisé en planification d'être lisible comme outil d'anticipation, voire de communication, que d'être fidèle aux processus territoriaux au prix d'une abstraction. Il est intéressant de noter que les priorités actuelles de développement des modèles LUTI semblent plus être centrées sur une meilleure intégration des nou-

velles technologies et une meilleur interaction avec les politique et la planification, par exemple via des interfaces de visualisation [Wee, 2015], mais en rien des problématiques de dynamiques territoriales incluant le réseau sur de plus longues échelles par exemple, ce qui confirme la portée et la logique autour de ce type de modèles.

Croissance du Réseau

La croissance de réseaux est pratiquée dans des entreprises de modélisation qui cherchent à expliquer de manière endogène, au sens de modèles génératifs, la croissance des réseaux de transport. Ils prennent généralement d'un point de vue *bottom-up*, i.e. en mettant en évidence des règles locales qui permettraient de reproduire la croissance du réseau sur de longues échelles de temps (souvent le réseau de rues). Les économistes ont proposés des modèles de ce type : [Zhang et Levinson, 2007] passe en revue la littérature en économie de transports sur la croissance des réseaux dans le contexte d'une théorie endogène de la croissance [Aghion et al., 1998], rappelant les trois aspects principalement traités par les économistes sur le sujet, qui sont la tarification routière, l'investissement en infrastructures et le régime de propriété, et propose finalement un modèle analytique combinant les trois. [Xie et Levinson, 2009c] propose une revue étendue de la modélisation de croissance des réseaux, en prenant en compte d'autres champs : la géographie des transports a développé très tôt des modèles basés sur des faits empiriques mais qui se sont concentrés sur reproduire la topologie plutôt que sur les mécanismes selon [Xie et Levinson, 2009c] ; les modèles statistiques sur des cas d'étude fournissent des conclusions très mitigées sur les relations causales entre croissance du réseau et demande (la croissance étant dans ce cas conditionnée aux données de demande) ; les économistes ont étudié la production d'infrastructure à la fois d'un point de vue microscopique et macroscopique, généralement non spatiaux ; la science des réseaux a produit des modèles jouet de croissance de réseau qui se basent sur des règles topologiques et structurelles plutôt que des règles se reposant sur des processus inspirés de faits réels. Nous donnons pour commencer des exemples d'études utilisant des concepts économiques ou géométriques pour modéliser la croissance de réseau. Les mécanismes induisant la croissance du réseau, sur le plan de la gouvernance ou économique, peuvent être très détaillés, comme [Levinson, Xie et Oca, 2012] qui se base sur des enquêtes qualitatives et des modèles statistiques calibrés sur des vraies données pour paramétrer un modèle de croissance de réseau. [Xie et Levinson, 2009b] compare l'influence relative des processus de croissance centralisés et décentralisés. [Levinson et Karamalaputi, 2003] procède à une étude empirique des déterminants de la croissance du réseau routier pour les Twin Cities, établissant que les variables basiques (longueur, changement dans l'accessibilité) ont le comportement attendu,

et qu'il existe une différence entre les niveaux d'investissement, impliquant que la croissance locale n'est pas affectée par les coûts, ce qui peut correspondre à une équité des territoires dans l'accessibilité minimale. [YERRA et LEVINSON, 2005] montre avec un modèle économique basé sur des processus auto-renforçants et incluant une règle d'investissement basée sur l'attribution du trafic, que des règles locales sont suffisantes pour faire émerger une hiérarchie du réseau routier à usage du sol fixé. Une modèle très similaire donnée par [LOUF, JENSEN et BARTHELEMY, 2013] avec des fonctions coûts-bénéfices plus simples obtient une conclusion similaire. Etant donné une distribution de noeuds (villes) dont la population suit une loi puissance, deux villes effectueront un lien si une fonction d'utilité coût-bénéfice combinant linéairement flux gravitaire potentiel (loi puissance de la distance) et coût de construction (linéaire de la distance) a une valeur positive. Ces hypothèses locales simples suffisent à faire émerger un réseau complexe et des transitions de phase en fonction du paramètre de poids relatif dans le coût, conduisant à l'apparition de la hiérarchie. Alors que ces modèles basés sur des processus cherchent à reproduire des motifs macroscopiques des réseaux (typiquement les lois d'échelle), les modèles d'optimisation géométrique cherchent à ressembler à des réseaux réels dans leur topologie. [BARTHÉLEMY et FLAMMINI, 2008] décrit un modèle basé sur une optimisation locale de l'énergie, mais ce modèle reste très abstrait et non validé. Le modèle de morphogenèse de [COURTAT, GLOAGUEN et DOUADY, 2011] qui utilise des potentiels locaux et des règles de connectivité, même s'il n'est pas calibré, semble reproduire de manière plus raisonnable des motifs réels des réseaux de rues. Un modèle très proche est décrit dans [RUI et al., 2013]. D'autres tentatives comme [DE LEON, FELSEN et WILENSKY, 2007 ; YAMINS, RASMUSSEN et FOGEL, 2003] sont plus proches de la modélisation procédurale [LECHNER et al., 2004 ; WATSON et al., 2008] et pour cette raison n'ont pas d'intérêt pour notre cas puisqu'ils peuvent difficilement être utilisés comme modèles explicatifs. La modélisation procédurale génère des structure à la manière des grammaires de forme, mais celle-ci se concentre généralement sur la reproduction fidèle de forme locale, sans tenir compte des propriétés macroscopiques émergentes. Les classifier comme modèles de morphogenèse n'est pas correct et correspond à une incompréhension des mécanismes du *Pattern Oriented Modeling* d'une part et de l'épistémologie de la Morphogenèse d'autre part (voir 6.1). Nous utiliserons ce type de modèle (mixtures d'exponentielles ou réseau par connexification) pour générer des données synthétiques initiales uniquement pour faire tourner d'autres modèles complexes (voir 3.2 et 6.3). Enfin, une approche originale et intéressante à la croissance des réseaux sont les réseaux biologiques. Ils appartiennent au champ de l'ingénierie morphogénétique dont DOURSAT est un pionnier, qui vise à concevoir des systèmes complexes artificiels inspirés de systèmes complexes

naturels et sur lesquels un contrôle des propriétés émergentes est possible [Doursat, Sayama et Michel, 2012]. Les *Machines Physarum*, qui sont des modèles d'une moisissure auto-organisée (*slime mould*) ont été prouvés comme résolvant de manière efficiente et par le bas des problèmes computationnellement lourds comme des problème de routage [Tero, Kobayashi et Nakagaki, 2006] ou des problèmes de navigation NP-complets comme le Problème du Voyageur de Commerce [Zhu et al., 2013a], ce qui est porteur de sens au regard des liens entre différents types de complexité développés en 3.3. Ils produisent des réseaux ayant des propriétés de coût-robustesse Pareto-efficientes [Tero et al., 2010] qui sont typique des propriétés empiriques des réseaux réels, et de plus relativement proches en forme de ceux-ci (sous certaines conditions, voir [Adamatzky et Jones, 2010]). Ce type de modèles peut être d'intérêt dans notre cas puisque les processus d'auto-renforcement basés sur les flots sont analogues aux mécanismes de renforcement de lien en économie des transports. Ce type d'heuristique a été testé pour générer le réseau ferré Français par [Mimeur, 2016], faisant un pont intéressant avec les modèles d'investissement de Levinson. Les critères de validation appliqués restent cependant limités, soit à un niveau inadapté aux faits stylisés étudiés (nombre d'intersection ou de branches) soit trop générales pouvant être produit par n'importe quel modèle (longueur totale et pourcentage de population desservie), et relèvent de critère de forme typique de la modélisation procédurale qui ne peuvent que difficilement rendre compte des dynamiques internes d'un système comme développé précédemment. De plus, prendre pour validation externe la production d'un réseau hiérarchique découle d'une exploration incomplète de la structure et du comportement du modèle, puisque celui-ci par ses mécanismes d'attachement préférentiel doit mécaniquement produire une hiérarchie.

2.1.3 *Modéliser la co-évolution*

Modélisation Hybride

Les modèles de simulation qui incluent un couplage des dynamiques de la croissance urbaine et du réseau de transport sont relativement rares, et pour la plupart au stade de modèles stylisés. Les efforts étant assez disparates et dans des domaines très variés, il est difficile de percevoir une unité dans ce type de modèle, si ce n'est l'abstraction de l'hypothèse d'interdépendance entre réseaux et caractéristiques du territoire dans le temps. Une généralisation du modèle d'optimisation locale géométrique décrit précédemment a été développé dans [Barthélemy et Flammini, 2009]. Comme pour le modèle de croissance de réseau routier dont il est l'extension, les mécanismes locaux n'ont pas de justification théorique ou thématique, et le modèle n'est de plus pas exploré et aucune connaissance géographique ne

peut en être tirée. [Levinson, Xie et Zhu, 2007] prend une approche économique plus intéressante du point de vue des processus de développement de réseau impliqués, similaire à un modèle à quatre étapes (génération de flux origine-destination basés sur la gravité, attribution du traffic par Equilibre Utilisateur Stochastique) qui inclut coût de transport et congestion, couplé avec un module d'investissement routier qui simule les revenus des péages pour les agents qui construisent, et un module d'évolution d'usage du sol qui met à jour les actifs et emplois par modélisation de choix discrets. Les expériences montrent que l'usage du sol et le réseau en co-évolution mène à des retroactions positives renforçant les hiérarchies, mais sont loin d'être satisfaisantes pour deux raisons : d'une part la topologie du réseau n'évolue pas à proprement parler puisque seules les capacités et les flux changent dans le réseau, ce qui signifie que des mécanismes plus complexes sur de plus longues échelles de temps ne sont pas pris en compte, et d'autre part les conclusions sont assez limitées puisque le comportement du modèle n'est pas connu, les analyses de sensibilité étant faites sur un petit nombre d'espaces unidimensionnels : les mécanismes exhaustifs restent ainsi inconnus comme seuls des cas particuliers sont donnés dans l'analyse de sensibilité. D'un autre point de vue, [Levinson et Chen, 2005] est aussi présenté comme un modèle de co-évolution mais correspond plus à une analyse statistique couplée puisqu'elle repose sur un modèle prédictif à chaîne de Markov. [Rui et Ban, 2011] décrit un modèle dans lequel le couplage entre usage du sol et la topologie du réseau est fait par un paradigme faible, l'usage du sol et l'accessibilité n'ayant pas de retroaction sur la topologie du réseau, le modèle d'usage du sol étant conditionné à la croissance du réseau autonome. Ce modèle est mis en perspective avec d'autres modèles d'usage du sol et de croissance de réseau dans [Rui, 2013]. [Achibet et al., 2014] décrit un modèle de co-évolution à une très petite échelle (échelle du bâtiment), dans lequel l'évolution du réseau et des bâtiments sont tous les deux régis par un agent commun (qui est influencé différemment par la topologie du réseau et la densité de population) ce qui implique une simplification trop grande des processus sous-jacents. Enfin, un modèle hybride simple exploré et appliqué à un exemple jouet de planification dans [Raimbault, Banos et Doursat, 2014], repose sur les mécanismes d'accès aux activités urbaines pour la croissance des établissements avec un réseau s'adaptant à la forme urbaine. Les règles pour la croissance du réseau sont trop simples pour capturer les processus qui nous intéressent, mais le modèle produit à une petite échelle une large gamme de formes urbaines qui reproduisent les motifs typiques des établissements humains. Ce modèle est s'inspire de [Moreno, Badariotti et Banos, 2012] pour ses mécanismes de base mais permet une génération de formes bien plus larges par la prise en compte des fonctions urbaines. A cette échelle, i.e. urbaine

ou métropolitaine, les mécanismes de localisation de population influencée par l'accessibilité couplés à des mécanismes de croissance de réseau optimisant certaines fonctions semblent être la règle pour ces modèles : de la même façon, [Wu et al., 2017] couple un CA de diffusion de population à un réseau optimisant un coût local dépendant de la géométrie et de la distribution de population. De manière conceptuelle, une certaine forme de couplage fort est opéré dans [Bigotte et al., 2010] qui par une approche de recherche opérationnelle propose un algorithme de design de réseau pour optimiser l'accessibilité aux services, prenant en compte à la fois la hiérarchie du réseau et celle des centres connecté. Enfin, le modèle proposé par [Blumenfeld-Lieberthal et Portugali, 2010] peut être vu comme une transition vers les approches de type système urbain, puisqu'il simule les migrations entre villes et la croissance du réseau induite par une rupture de potentiel lorsque les détours sont trop grands. A une échelle macroscopique et également plus proche de la modélisation de système urbains que nous développerons dans la section suivante, [Baptiste, 1999] propose de coupler le modèle de croissance urbaine basé sur les migrations (introduit par l'application de la synergétique au système de ville par Sanders dans [Sanders, 1992]) avec un mécanisme d'auto-renforcement pour le réseau routier sans modification topologique (retroaction positive par seuils du différentiel flux-capacité sur la capacité). Sa dernière version est présentée par [Baptiste, 2010]. Guère de conclusions générales ne peuvent cependant être tirées de ce travail, outre que ce couplage permet de faire émerger une configuration hiérarchique (mais on sait par ailleurs que des modèles plus simples, un attachement préférentiel uniquement par exemple, permettent de reproduire ce fait stylisé) et que l'ajout du réseau produit un espace moins hiérarchique, permettant à des villes moyennes de bénéficier de la rétroaction du réseau de transport.

Modélisation de Systèmes Urbains

Une approche relativement proche des précédentes, mais ayant des caractéristiques propres, est celle de la modélisation intégrée des systèmes de villes. Dans la continuité des modèles Simpop pour modéliser les systèmes de villes, Schmitt décrit dans [Schmitt, 2014] le modèle SimpopNet qui vise à précisément intégrer les processus de co-évolution dans les systèmes de villes à longue échelle temporelle, typiquement par des règles pour un développement hiérarchique du réseau comme fonction des dynamiques des villes, couplées à celles-ci qui dépendent de la topologie du réseau. Malheureusement le modèle n'a pas été exploré ni étudié de manière plus approfondie, et de plus est resté au niveau de modèle jouet. Cottineau propose une croissance endogène des réseaux de transport comme la dernière brique de construction de ses productions Marius [Cottineau, 2014] mais cela reste à un niveau conceptuel puisque cette brique n'a pas

encore été spécifiée ni implémentée. Il n'existe à notre connaissance pas de modèle empirique ou appliqué à un cas concret se basant sur une approche de la co-évolution par les systèmes urbains vus par la Théorie Evolutive des Villes. Nous nous positionnerons particulièrement dans cette lignée de recherche dans cette thèse, vu l'importance que prendra la Théorie Evolutive dans notre démarche Théorique et de Modélisation comme nous le détaillerons par la suite. L'ensemble des briques est nécessaire pour comprendre les implications de ce positionnement, mais le lecteur pressé pourra directement consulter le chapitre 9 pour une synthèse des implications théoriques à différents niveaux d'abstraction. Typiquement, les hypothèses épistémologiques fondamentales tel le rôle des relations et de la configuration spatiales, ou la présence d'un équilibre - nous considérons les systèmes urbains comme des systèmes complexes adaptatifs, auto-organisés loin de l'équilibre, sont typiques de cette approche si on les considère conjointement. On voit bien l'opposition aux principes épistémologiques de l'économie géographique : [Fujita, Krugman et Mori, 1999] introduit par exemple un modèle évolutionnaire capable de reproduire une hiérarchie urbaine et une organisation typique de la Théorie des Places Centrales, mais repose toujours sur la notion d'équilibres successifs, et surtout considère un modèle "à-la-krugman" c'est à dire un espace à une dimension homogène. Cette approche peut être instructive sur les processus économiques en eux-mêmes mais aucunement sur les processus géographiques, qui incluent le déroulement des processus économiques dans l'espace géographique dans lequel les particularités sont essentielles. Notre travail s'attellera à montrer dans quelle mesure cette structure de l'espace peut être importante et également explicative, puisque les réseaux , et encore plus les réseaux physiques induisent des processus dépendants au chemin spatio-temporel et donc sensibles au singularités locales et propices aux bifurcations induites par la combinaison de celles-ci et de processus à d'autres échelles (par exemple la centralité induisant un flux).

Co-évolution

Après cet aperçu de la littérature, incluant différents degrés de couplage entre les composantes des réseaux et territoires, nous sommes en mesure de préciser ce que nous entendrons par *modéliser la co-évolution*. La vision donnée ici a un but opérationnel, puisqu'il ne nous paraissait pas pertinent de donner d'emblée une vision trop théorique et abstraite (qui sera développée en 9.1). En Géographie Economique, la notion de co-évolution a également été mobilisée, notamment dans sa branche évolutionnaire. Ainsi, [Wal et Boschma, 2011] introduit un cadre conceptuel pour permettre de concilier nature évolutionnaire des firmes, théorie des clusters et réseaux de connaissance, dans lequel la co-évolution entre réseaux et firmes est centrale,

et qui est définie comme une causalité circulaire entre différentes caractéristiques de ces sous-systèmes. L'idée d'entités évolutionnaire en économie est difficilement compatible avec le courant néoclassique mainstream, mais trouve un écho de plus en plus pertinent [Nelson et Winter, 2009]. Pour la géographie, les travaux les plus proches empiriquement et théoriquement des notions de co-évolution sont étroitement liés à la Théorie Evolutive des Villes. Il n'est pas évident de tracer dans la littérature à quel moment la notion s'est cristallisée, mais il est évident qu'elle était présente dès les fondements de la théorie comme le rappelle Denise Pumain (voir D.5) : le système complexe adaptatif est composé par des entités en dépendance causales fortes. Les premiers modèles incluent bien cette vision de manière implicite, mais la co-évolution n'est pas appuyée explicitement ou définie précisément, en termes qui seraient quantifiables ou identifiables structurellement. [Paulus, 2004] a amené des évidences empiriques de mécanismes de co-évolution par l'étude de l'évolution des profils économiques des villes françaises. L'interprétation utilisée par [Schmitt, 2014] repose sur une lecture par la Théorie Evolutive, mais reste très floue au delà d'une lecture des systèmes de villes comme entités fortement interdépendantes. Or l'interdépendance est une notion aussi lâche que le fameux "tout interagit avec tout", c'est à dire qu'elle est particulièrement creuse si elle n'est pas quantifiée. Elle permet comme prémisse épistémologique de considérer certaines ontologies et certaines démarches de modélisation, mais ne permet pas de comprendre finement la structure et les processus d'un système. Par exemple, étant donné un réseau topologique d'interaction entre entités et des motifs temporels de propagation correspondants, on peut se demander quels sont les motifs de corrélations statiques et dynamiques correspondants, s'il existe des causalités et à quelles échelles. Il existe en pratique une infinité de "régimes" de co-évolution possibles, liés à la structure du réseau écologique de la niche correspondante si on interprète celle-ci de cette façon [Holland, 2012]. L'idée de diffusion hiérarchique de l'innovation dans la théorie évolutive capture par exemple qualitativement certains de ces aspects, mais la quantification des régimes correspondants et donc de la co-évolution reste une question ouverte. L'une de nos contributions principales, qui aboutira comme produit des efforts empiriques et de modélisation, sous forme théorique en 9.1, sera de clarifier cette notion et d'en donner une définition précise. A ce point, l'état de l'art fait ci-dessus témoigne d'une faiblesse de la littérature dans le domaine du couplage fort entre évolution des territoires et croissance des réseau, vu la faible épaisseur et la disparité des travaux revus. Les lacunes à combler sur ce point seraient donc liées à l'introduction de modèles fortement couplés dans le temps plus ou moins multi-processus et multi-échelles, pour lesquels une partie des modèles décrits ci-dessus sont précurseurs.

⋆ ⋆

⋆

2.2 UNE APPROCHE EPISTÉMOLOGIQUE

Un corolaire de la matière thématique introduite en chapitre 1 est le besoin d'une compréhension des disciplines impliquées elles-même pour être en mesure de construire des modèles hétérogènes intégrés. Les possibilités de couplage et d'intégration sont hautement déterminées par les approches existantes et les lacunes correspondantes qui ont été exposées dans la section précédente 2.1. Cela implique une étude épistémologique avancée dans chaque champ, que nous proposons de mener de manière quantitative et systématique. Ce choix délibéré pourrait occulter des considérations épistémologiques élaborées mais suit notre objectif d'investigations préliminaires pour la construction de modèles, en révélant potentiellement des directions de recherche.

Nous décrivons et explorons d'abord un algorithme de revue systématique algorithmique, qui reconstruit des corpus de références par une extraction sémantique itérative. Nous procédons ensuite à une analyse de réseaux, couplant réseau de citation et réseau sémantique, pour préciser les contours des disciplines impliquées. Nous suggérons finalement des possibles extensions vers de l'apprentissage non-supervisé et la fouille de texte complets pour une extraction automatique de la structure de modèles par exemple.

2.2.1 *Revue Systématique Algorithmique*

Une étude bibliographique étendue suggère une rareté des modèles quantitatifs de simulation qui intègrent à la fois la croissance urbaine et la croissance des réseaux. Cette absence pourrait être due aux intérêts divergents des disciplines concernées qui induiraient un manque de communication. Nous proposons de procéder à une revue de la littérature systématique et algorithmique pour donner des éléments de réponse quantitatifs à cette question. Un algorithme itératif formel pour construire des corpus de références à partir de mots-clés initiaux, basé sur l'analyse textuelle, est développé et mis en oeuvre. Nous étudions ses propriétés de convergence et procédons à une analyse de sensibilité. Nous l'appliquons ensuite à des requêtes représentatives de notre question spécifique, pour lesquelles les résultats tendent à confirmer l'hypothèse d'isolation des disciplines.

En recherche de modèles de co-évolution

Comme développé en 1.1, les réseaux de transport et l'usage du sol urbain sont connus pour être des composantes au couplage complexe au sein des systèmes urbains, à différentes échelles [Bretagnolle, Pumain et Vacchiani-Marcuzzo, 2009]. Une approche commune est de les considérer comme étant en co-évolution, tout en évitant les interprétations trompeuses comme le mythe des effets structurants

des infrastructures de transport [Offner, 1993]. Une question qui se présente rapidement est l'existence de modèles endogénéisant cette co-évolution, i.e. prenant en compte simultanément la croissance urbaine et celle du réseau. Nous essayons d'y répondre par une revue systématique algorithmique. Nous proposons dans cette section de développer cette approche en formalisant l'algorithme, dont les résultats sont ensuite présentés et discutés.

Modéliser les Interactions entre croissance urbaine et croissance des réseaux

Nous avons revu selon divers point de vue les efforts de modélisation des interactions entre territoires et réseaux dans la section précédente 2.1. Cet état de l'art nous suggère fortement des domaines relativement cloisonnés et s'intéressant à des problématiques différentes.

Analyse Bibliométrique

Avec l'avènement des nouveaux moyens techniques et des nouvelles sources de données, la revue de littérature classique tend à se coupler à des revues automatiques. Des techniques de revue systématique ont été développées, des revues qualitatives aux meta-analyses quantitatives qui permettent de produire des nouveaux résultats par combinaison d'études existantes [Rucker, 2012]. Passer sous silence certaines références peut même être considéré comme une erreur scientifique dans le contexte de l'émergence des systèmes d'information qui par l'accès plus aisé à l'information rend difficilement justifiable l'omission de références clés [Lissack, 2013]. Nous proposons de tirer parti de telles techniques pour traiter notre problème. En effet, l'observation de la bibliographie obtenue dans la section précédente soulève une hypothèse. On peut postuler sans risques à partir de la revue précédente 2.1 Il semble clair que toutes les briques sont présentes pour l'existence de modèles co-évolutifs mais des questionnements et objectifs différents semblent la stopper. Comme montré par [Commenges, 2013b] pour le concept de mobilité, pour lequel un "petit monde d'acteurs" relativement fermé, en l'occurence les corsards des Ponts, a inventé une notion ad hoc, utilisant des modèles sans connaissance préalable d'un contexte scientifique plus général. On pourrait se trouver dans un cas similaire pour le type de modèles auxquels on s'intéresse. Des interactions restreintes entre des champs scientifiques travaillant sur les mêmes objets mais avec des objectifs et contextes divergents, et à des échelles différentes, pourrait être à l'origine de l'absence de modèles co-évolutifs. Tandis que la majorité des études en bibliométrie se reposent sur les réseaux de citation [Newman, 2013] ou les réseaux de co-auteurs [Sarigöl et al., 2014], nous proposons d'utiliser un paradigme moins exploré, basé sur l'analyse textuelle, introduit par [Chavalarias et Cointet, 2013],

qui obtient une cartographie dynamique des disciplines scientifiques en se basant sur leur contenu sémantique. Nous postulons que cette couche supplémentaire d'information apporte un information complémentaire, nécessaire pour appréhender la diversité des domaines. La méthode est particulièrement adaptée pour notre étude puisque nous voulons comprendre la structure du contenu des recherches sur le sujet. Nous appliquons une approche algorithmique décrite par la suite. L'algorithme procède par itérations pour obtenir un corpus stabilisé à partir de mots-clés initiaux, reconstruisant l'horizon sémantique scientifique autour d'un sujet donné.

DESCRIPTION DE L'ALGORITHME Soit A un alphabet (un ensemble arbitraire de symboles), A^* les mots correspondants et $T = \cup_{k\in\mathbb{N}} A^{*^k}$ les textes de longueur finie sur celui-ci. Ce qu'on nomme une référence est pour l'algorithme un enregistrement avec des champs textuels représentant le titre, le résumé et les mots-clés. L'ensemble de références à l'itération n sera noté $\mathcal{C} \subset T^3$: il s'agit d'un sous-ensemble de triplets de textes. Nous supposons l'existence d'un ensemble de mots-clés $\mathcal{K}_n$, les mots-clés initiaux étant $\mathcal{K}_0$, spécifiés par l'utilisateur[1]. Une itération procède de la manière suivante :

1. Un corpus intermédiaire brut $\mathcal{R}_n$ est obtenu par une requête à un catalogue[2] auquel on fourni les mots-clés précédents $\mathcal{K}_{n-1}$.
2. Le corpus total est actualisé par $\mathcal{C}_n = \mathcal{C}_{n-1} \cup \mathcal{R}_n$.
3. Les nouveaux mot-clés $\mathcal{K}_n$ sont extraits du corpus par Traitement du Language Naturel (NLP), étant donné un paramètre fixé N_k donnant le nombre de mot-clés.

L'algorithme termine quand la taille du corpus devient stable ou quand un nombre maximal d'itérations défini par l'utilisateur est atteint. La figure 1 synthétise le processus général.

RÉSULTATS Les détails précis concernant l'implémentation de l'algorithme ainsi qu'une analyse de sensibilité pour vérifier la convergence sur un échantillon de requêtes initiales (typiques des champs étudiés) sont donnés en Appendice A.2. Lorsque l'algorithme a été partiellement validé par cette analyse, nous l'appliquons à notre question. Nous partons de cinq différentes requêtes initiales qui ont été

1 On pourrait également partir d'un corpus $\mathcal{C}_0$, mais il s'agit plutôt de l'esprit de la méthodologie présentée dans la sous-section suivante. Nous nous en tiendrons ici pour cette exploration préliminaire en assumant le caractère arbitraire forcément biaisé de cette spécification.

2 La dépendance au catalogue devant sûrement introduire un biais que nous ne pouvons contrôler, une analyse de sensibilité ou le croisement de divers catalogues étant hors de propos pour cette analyse exploratoire.

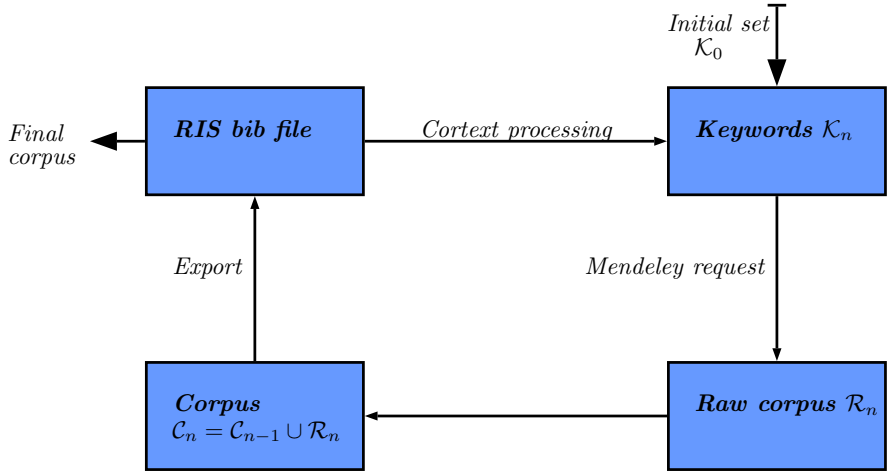


FIGURE 1 : Architecture globale de l'algorithme, incluant des détails d'implémentation : la requête au catalogue est faite via l'API Mendeley ; les corpus finaux sont sous forme de fichiers RIS.

manuellement extraites des divers domaines identifiés dans la bibliographie (qui sont "city system network", "land use transport interaction", "network urban modeling", "population density transport", "transportation network urban growth")[3]. Nous prenons l'hypothèse la plus faible pour le paramètre $N_k = 100$ (plus N_k est grand, plus les domaines atteints devraient être moins restreints et donc plus des résultats de distance seront significatifs). Après avoir construit les corpus, nous étudions leur cohérence lexicale comme un indicateur de réponse à notre question initiale. De grande distances devraient confirmer l'hypothèse formulée ci-dessus, i.e. que des disciplines auto-centrées pourraient être à l'origine d'un manque d'intérêt pour des modèles co-évolutifs. La table ?? montre les valeurs de la proximité lexicale relative, qui est significativement basse sachant que les chiffres peuvent directement s'interpréter comme une proportion de mots en co-occurrence, ce qui tend à confirmer notre hypothèse. Pour être plus précis tout de même, il faudrait un modèle nul avec des corpus aléatoires par exemple, ce qui pourrait faire l'objet de développements futurs.

Les développements possibles incluent la construction de réseaux de citation via un accès automatique à Google Scholar qui fournit les citations entrantes. La confrontation des coefficients inter-clusters pour le réseau de citations entre les différents corpus avec la cohérence lexicale est un aspect clé d'une validation approfondie des résultats.

L'absence peu explicable a priori de modèles qui simulent la co-évolution des réseaux de transport et de l'usage du sol urbain, qui se confirme à première vue par un état de l'art couvrant des do-

3 Ce choix est arbitraire, cette étude étant préliminaire on admet de travailler potentiellement sur des échantillons. Par exemple, l'utilisation de "co-evolution" n'est pas concluante car trop peu d'articles utilisent cette formulation.

Corpuses	1	2	3	4	5
1 (W=3789)	1	0	0.0719	0.0078	0.0724
2 (W=5180)	0	1	0.0338	0	0.0125
3 (W=3757)	0.0719	0.0338	1	0.0100	0.1729
4 (W=3551)	0.0078	0	0.0100	1	0.0333
5 (W=8338)	0.0724	0.0125	0.1729	0.0333	1

TABLE 1 : Matrice symétrique des proximités lexicales entre les corpus finaux, définies comme la somme des co-occurrences totale de mots-clés finaux entre corpus, normalisé par le nombre de mots-clés finaux (100). La taille des corpus finaux est donnée par *W*. Les valeurs obtenues pour les proximités sont considérablement faibles, ce qui confirme que les corpus sont éloignés de manière significative (voir texte).

maines disparates, pourrait être due à l'absence de communication entre les disciplines scientifiques étudiant différents aspects du problème. D'autres explications possibles qui en sont proches peuvent par exemple être le manque de cas d'application concrets de tels modèles vu les échelles temporelles mises en jeu et donc l'absence de financement propre - ce qui n'est pas si loin de l'absence d'une discipline y consacrant certains de ses objets. Cette question des portées et des échelles des modèles fera l'objet de la meta-analyse à la section suivante 2.3. Ainsi, nous ici avons proposé une méthode algorithmique pour donner des éléments de réponse par l'extraction de corpus basée sur l'analyse textuelle. Les premiers résultats numériques semblent confirmer l'hypothèse. Cependant, une telle analyse quantitative ne doit pas être considérée seule, mais devrait plutôt venir comme soutien à des études qualitatives qui peuvent être l'objet de développements futurs, comme celle menée dans [COMMENGES, 2013b], dans laquelle des questionnaires avec des acteurs historiques fournit des informations extrêmement pertinentes.

2.2.2 *Bibliométrie Indirecte par Analyse de Réseaux Complexes*

Comme décrit précédemment, l'analyse sémantique des corpus finaux ne contient pas la totalité de l'information sur les liens entre disciplines ni sur les motifs de propagation de la connaissance scientifique comme ceux contenus dans les réseaux de citations par exemple. De plus, la collection des données dans l'algorithme précédent est sujette à convergence vers des thèmes relativement auto-cohérents de par la structure propre de la méthode. On pourrait obtenir plus d'information sur les motifs sociaux de choix ontologiques pour la modélisation en étudiant les communautés dans des réseaux plus larges, ce qui correspondrait plus à des disciplines (ou des sous-disciplines

selon le niveau de granularité). Nous proposons de reconstruire les disciplines autour de notre thématique, pour obtenir une vue plus précise de l'interdisciplinarité et du paysage scientifique sur notre sujet.

Contexte

La majorité des disciplines scientifiques présentent un besoin fort en interdisciplinarité et approches transversales, comme illustré par exemple par l'édition spéciale récente de Nature sur le sujet (["Interdisciplinarity, Nature Special Issue"]), pour diverses raisons qui peuvent inclure le développement de champs intégrés verticalement conjointement aux questions horizontales comme détaillé dans la feuille de route des Systèmes Complexes ([Bourgine, Chavalarias et al., 2009]). Les débats courants sur la nature exacte de l'interdisciplinarité sont bien sûr nombreux (d'autres termes existent comme transdisciplinarité ou cross-disciplinarité), et celle-ci dépend en fait des domaines impliqués : des disciplines hybrides apparues récemment (voir par exemples celles soulignées par [Bais, 2010] comme l'astrobiologie) sont une bonne illustration du cas où les intrications sont très fortes, tandis que des champs plus mou comme "l'urbanisme" qui n'ont pas de définition précise montrent comment l'intégration horizontale est nécessaire et comment de la connaissance transversale peut être produite (menant à des possibles malentendus lorsque récemment introduite trop brutalement à des physiciens comme montré par [Dupuy et Benguigui, 2015]). Cette question se transfère naturellement au champ de la communication scientifique : quelles sont les alternatives correspondantes pour une dissémination efficace de la connaissance ? Des éléments de réponse à une question si générale impliquent, dans une perspective evidence-based, des mesures quantitatives de l'interdisciplinarité, qui font partie d'une approche multidimensionnelle de l'étude de la science, en quelque sorte "au-delà de la bibliométrie" [Cronin et Sugimoto, 2014].

Les méthodes potentielles pour des entrées quantitatives en épistémologie sont nombreuses. En utilisant les caractéristiques des réseaux de citation, un bon pouvoir prédictif pour les motifs de citation est par exemple obtenu par [Newman, 2013]. Les réseaux de co-auteurs peuvent également être utilisés pour des modèles prédictifs [Sarigöl et al., 2014]. Une approche multi-couches a récemment été proposée par [Omodei, De Domenico et Arenas, 2016], utilisant des réseaux bipartites des papiers et des chercheurs, dans le but de produire des mesures d'interdisciplinarité. Les disciplines peuvent être stratifiées en couches pour révéler des communautés entre elles et ainsi des motifs de collaboration [Battiston et al., 2015]. Les réseaux de mots-clés sont utilisés dans d'autres champs comme l'économie de l'innovation : par exemple, [Choi et Hwang, 2014] introduit une méthode pour identifier les opportunités technologiques en détectant des mots-

clés importants au sens des mesures topologiques. [SHIBATA et al., 2008] utilise l'analyse topologique du réseau de citations pour détecter des fronts de recherche émergents.

L'approche développée ici couple exploration et analyse de réseau de citation avec analyse textuelle, dans le but de cartographier le paysage scientifique dans le voisinage d'un corpus donné. Le contexte est particulièrement intéressant pour la méthodologie développée. Premièrement, le sujet étudié est très large et par essence interdisciplinaire. Deuxièmement, les données bibliographiques sont difficiles à obtenir, soulevant la question de comment la perception d'un horizon scientifique peut être déterminée par les acteurs de la dissémination et donc loin d'être objective, rendant les solutions techniques comme celle développée ici en conséquence des outils cruciaux pour une science ouverte et neutre. Notre approche combine une analyse des communautés sémantiques (comme fait dans [PALCHYKOV et al., 2016] pour les articles en physique mais sans extraction des mots-clés, ou par [GURCIULLO et al., 2015] pour un analyse des réseaux sémantiques de débats politiques) avec celle du réseau de citations pour extraire par exemple des mesures d'interdisciplinarité. Notre contribution se démarque des travaux précédents quantifiant l'interdisciplinarité puisqu'elle ne suppose pas de domaines a priori ou une classification des références considérées, mais reconstruit par le bas les champs via l'information sémantique endogène. [NICHOLS, 2014] introduit une approche similaire, utilisant le modèle d'extraction de thématiques *Latent Dirichlet Allocation* pour caractériser l'interdisciplinarité de récompenses dans des sciences précises. [LARIVIÈRE et GINGRAS, 2014] quantifie l'interdisciplinarité sur une longue période temporelle en étudiant l'étendue de la bibliographie des publications.

Données

Notre approche implique des spécifications pour le jeu de données utilisé, à savoir : (i) couvrir un voisinage conséquent du corpus étudié dans le réseau de citation afin d'avoir une vue la moins biaisée possible du paysage scientifique ; (ii) avoir au moins une description textuelle pour chaque noeud. Pour cela, nous rassemblons et compilons les données de sources hétérogènes en utilisant une architecture et implémentation spécifiques, décrites en Appendice B.6. Pour simplifier, nous dénommons *référence* toute production scientifique standard[4] qui peut être citée par une autre (articles de journaux, livre, chapitre de livre, article d'actes, communication, etc.) et contient des informations de base (titre, résumé, auteurs, année de publication). Nous travaillons par la suite sur le réseau des références. Il est important de noter qu'une contribution fondamentale de cette partie consiste en la construction de jeux de données hybrides à partir de sources hé-

4 ce qui est bien sûr sujet à débat, voir nos discussion en ouverture sur l'évolution des modes de communication scientifique

térogènes, et les développement des outils associés qui peuvent être réutilisés et améliorés pour des applications similaires.

CORPUS INITIAL Notre corpus initial est construit à partir de l'état de l'art établi en 2.1. Sa composition complète est donnée en Appendice A.2. Celui-ci est pris de taille raisonnable, mais les méthodes utilisées ici ont été développées sur des données massives, pour les brevets par exemple [BERGEAUD, POTIRON et RAIMBAULT, 2017a].

DONNÉES DE CITATION Le réseau de citations est reconstruit à partir de `Google Scholar` qui est souvent l'unique source des citations entrantes [NORUZI, 2005] puisqu'en science humaines les ouvrages ne sont pas systématiquement référencés par les bases fournissant des services (payants) comme le réseau de citation.[5] Nous sommes conscients des biais possibles de l'utilisation de cette source unique (voir par exemple [BOHANNON, 2014])[6], mais ces critiques sont plutôt dirigées vers les résultats de recherche plutôt que les comptes de citations. Nous récoltons ainsi les références *citantes* à profondeur deux, c'est à dire les références citant le corpus initial et celles citant celles-ci. Le réseau obtenu contient $V = 9462$ références correspondant à $E = 12004$ liens de citation. Concernant les langues, l'anglais représente 87% du corpus, le français 6%, l'espagnol 3%, l'allemand 1%, complété par des langues comme le mandarin pouvant être indéfinies (la détection de celui-ci étant peu fiable). Le corpus n'est pas très international (contrairement par exemple au thème de la croissance urbaine, étudié pour le développement thématique sur les liens entre économie et géographie développé en C.1).

DONNÉES TEXTUELLES Pour mener l'analyse sémantique, une description suffisamment conséquente est nécessaire. Nous collectons pour cela les résumés pour le réseau précédent. Ceux-ci sont disponibles pour environ un tiers des références, donnant $V = 3510$ noeuds avec description textuelle.

Résultats

RÉSEAU DE CITATIONS Des statistiques basiques pour le réseau de citation donnent déjà des informations intéressantes. Le réseau a un degré moyen de $\bar{d} = 2.53$ et une densité de $\gamma = 0.0013$. Le degré entrant moyen (qui peut être interprété comme un facteur d'impact stationnaire) est de 1.26, ce qui est relativement élevé pour des sciences humaines. Il est important de noter sa connexité faible, ce qui signifie que les domaines initiaux ne sont pas en isolation totale : les références initiales sont partagées à un degré minimal par les dif-

5 Par exemple, le journal Cybergeo n'est indexé dans le *Web of Science* que depuis mai 2016, suite à des négociations ardues et non sans contrepartie.

6 ou `http:iscpif.frblog201602the-strange-arithmetic-of-google-scholars`

férents domaines. Nous travaillons sur la suite sur le sous-réseau des noeuds comprenant au moins deux liens, pour extraire le coeur de la structure du réseau et se débarrasser de l'effet "grappe". De plus, le réseau est nécessairement complet entre ces noeuds puisqu'on est remonté au deuxième niveau. Nous procédons à une détection de communautés par l'algorithme de Louvain, sur le réseau non-dirigé correspondant. On obtient 13 communautés, de modularité dirigée 0.66, extrêmement significative en comparaison à une estimation par bootstrap de la même mesure sur le graphe aléatoirement rebranché qui donne une modularité de 0.0005 ± 0.0051 sur $N = 100$ répétitions. Les communautés font sens de manière thématique, puisqu'on retrouve pour les plus grosses les domaines suivants : LUTI (18% du réseau), Géographie Urbaine et des Transports (16%), Planification des infrastructures (12%), Planification intégrée - TOD (6%), Réseaux Spatiaux (17%), Etudes d'accessibilité (18%). La Fig. 2 permet de visualiser les relations de ces domaines. Il est intéressant d'observer que les travaux des économistes et des physiciens dans le domaine tombent dans la même catégorie d'étude des *Spatial Networks*. En effet, la littérature citée par les physiciens comporte souvent plus d'ouvrage en économie qu'en géographie, tandis que les économistes utilisent des techniques d'analyse de réseau. Ensuite, le planning, l'accessibilité, les LUTI et le TOD sont très proches mais se distinguent dans leur spécificités : le fait qu'ils apparaissent dans des communautés séparées est un résultat en lui-même témoignant d'une certaine séparation. Ceux-ci font le pont entre les approches Réseaux spatiaux et les approches géographiques, qui comportent une partie importante de sciences politiques par exemple. Les liens entre physique et géographie restent très faibles. Ce panorama dépend bien sûr du corpus initial, mais nous permet de mieux comprendre le contexte de celui-ci dans son environnement disciplinaire.

COMMUNAUTÉS SÉMANTIQUES L'extraction des mots-clés est faite suivant une heuristique inspirée de [CHAVALARIAS et COINTET, 2013]. La description complète de la méthode et de son implémentation est donnée en Appendice B.6. Elle se base sur les relations au second ordre entre les entités sémantiques, qui sont des *n-grams*, c'est à dire des mots-clés multiples pouvant avoir une longueur jusqu'à 3. Celles-ci sont estimées via la matrice de co-occurrence, dont les propriétés statistiques fournissent une mesure de déviation à des co-occurrences uniformes, qui est utilisée pour juger la pertinence des mots-clés. Sélectionnant un nombre fixe de mots-clés pertinents $K_W = 10000$, nous pouvons ensuite construire un réseau pondéré par les co-occurrences.

La topologie du réseau brut ne permet pas l'extraction claire de communautés, en particulier à cause de hubs qui correspondent à des termes fréquents commun à de nombreux champs (e.g. `model`, `space`). Nous faisons l'hypothèse que ces termes à fort degré ne portent pas

FIGURE 2 : **Réseau de citations.** Nous visualisons les références ayant au moins deux liens, par un algorithme de force-atlas. Les couleurs donnent les communautés décrites dans le texte. En orange, bleu, turquoise : géographie urbaine, géographie des transports, sciences politiques ; en rose, noir, vert : planning, accessibilité, LUTI ; en violet : réseaux spatiaux (physique et économie).

d'information particulière sur des classes données et peuvent ainsi être filtrés étant donné un seuil de degré maximal k_{max} (on s'intéresse alors à ce qui fait la spécificité de chaque domaine). De la même manière, les liens avec un poids faibles sont considérés comme du bruit et filtrés selon un seuil de poids minimal θ_w. La méthode générique permet de plus une filtration préliminaire des mot-clés, complémentaire à la filtration topologique, par fréquence d'apparition dans les documents $[f_{min}, f_{max}]$, à laquelle les résultats ne sont pas sensibles dans notre cas. L'analyse de sensibilité des caractéristiques du réseau filtré, notamment de sa taille, modularité et structure des communautés, est donnée en A.2. Nous choisissons des valeurs de paramètres permettant une optimisation multi-objectif entre modularité et taille du réseau, $\theta_w = 10, k_{max} = 500$, par le choix d'un point com-

promis sur un front de Pareto, qui donne un réseau sémantique de taille ($V = 7063, E = 48952$). Celui-ci est visualisé en Appendice A.2.

Nous récupérons ensuite les communautés dans le réseau par un clustering de Louvain standard sur le réseau filtré optimal. On obtient 20 communautés pour une modularité de 0.58. Celles-ci sont examinées à la main pour être nommées, les techniques de désignation automatique [YANG et al., 2000] n'étant pas assez élaborées et ne font pas la distinction implicite entre champs thématiques et méthodologiques par exemple (en fait entre les domaines de connaissance, voir 9.3) qui est une dimension supplémentaire que nous ne traitons pas ici, mais nécessaire pour avoir des désignations parlantes. Les communautés sont décrites en Table **??**. On voit tout de suite la complémentarité avec l'approche par citation, puisque se dégagent ici à la fois des sujet d'étude (High Speed Rail, Maritime Networks), des domaines et méthodes (Networks, Remote Sensing, Mobility Data Mining), des domaines thématiques (Policy), des méthodes pures (Agent-based Modeling, Measuring). Ainsi, une référence peut mobiliser plusieurs de ces communautés. On a de plus une granularité plus fine de l'information. L'effet du langage est puissant puisque la géographie française se distingue en une catégorie séparée (des analyses poussées pourraient être envisagées pour mieux comprendre le phénomène et en tirer parti : sous-communautés, reconstruction d'un réseau spécifique, études par traduction ; mais celles-ci sont hors de propos dans cette étude exploratoire). On constante l'importance des réseaux, des problématiques de sciences politiques et socio-économiques. Nous mobiliserons la première catégorie dans la plupart des modèles développés, mais en gardant en tête l'importance des problématiques liées à la gouvernance, nous réaliserons un travail spécifique en 8.3.

MESURES D'INTERDISCIPLINARITÉ La distribution des mots clés dans les communautés permettent de définir une mesure d'interdisciplinarité au niveau de l'article. La combinaison des couches de citation et sémantique dans l'hyperréseau fournit des mesures d'interdisciplinarité au second ordre (motifs sémantiques des cités ou des citants), que nous n'utiliserons pas ici à cause de la taille modeste du réseau de citation (voir B.6 et **??**). Plus précisément, une référence i peut être vue comme un vecteur de probabilités sur les classes sémantiques j, qu'on notera sous forme matricielle $\mathbf{P} = (p_{ij})$. Celles-ci sont estimées simplement par les proportions de mots-clés classifiés dans chaque classe pour la référence. Une mesure classique d'interdisciplinarité [BERGEAUD, POTIRON et RAIMBAULT, 2017a] est alors $I_i = 1 - \sum_j p_{ij}^2$. Soit $\mathbf{A}$ la matrice d'adjacence du réseau de citation, et soit $\mathbf{I}_k$ les matrices de selection des lignes correspondants à la classe k de la classification de citation : $Id \cdot \mathbb{1}_{c(i)=k}$, telle que $I_k \cdot A \cdot I_{k'}$ donne exactement les citations de k vers k'. La proximité de citation entre les communautés de citation est alors définie par

TABLE 2 : **Description des communautés sémantiques.** On donne leur taille, leur proportion en quantité de mots-clés cumulés sur l'ensemble du corpus, et des mots-clés représentatifs sélectionnés par degré maximal.

Name	Size	Weight	Keywords
Networks	820	13.57%	social network, spatial network, resili
Policy	700	11.8%	actor, decision-mak, societi
Socio-economic	793	11.6%	neighborhood, incom, live
High Speed Rail	476	7.14%	high-spe, corridor, hsr
French Geography	210	6.08%	système, développement, territoire
Education	374	5.43%	school, student, collabor
Climate Change	411	5.42%	mitig, carbon, consumpt
Remote Sensing	405	4.65%	classif, detect, cover
Sustainable Transport	370	4.38%	sustain urban, travel demand, activity-bas
Traffic	368	4.23%	traffic congest, cbd, capit
Maritime Networks	402	4.2%	govern model, seaport, port author
Environment	289	3.79%	ecosystem servic, regul, settlement
Accessibility	260	3.23%	access measur, transport access, urban growth
Agent-based Modeling	192	3.18%	agent-bas, spread, heterogen
Transportation planning	192	3.18%	transport project, option, cba
Mobility Data Mining	168	2.49%	human mobil, movement, mobil phone
Health Geography	196	2.49%	healthcar, inequ, exclus
Freight and Logistics	239	2.06%	freight transport, citi logist, modal
Spanish Geography	106	1.26%	movilidad urbana, criteria, para
Measuring	166	1.0%	score, sampl, metric

$c_{kk'} = \sum \mathbf{I}_k \cdot \mathbf{A} \cdot \mathbf{I}_{k'} / \sum \mathbf{I}_k \cdot \mathbf{A}$. On définit la proximité sémantique en définissant une matrice de distance entre références par $\mathbf{D} = d_{ii'} = \sqrt{\frac{1}{2}\sum (p_{ij} - pi'j)^2}$ puis la proximité sémantique par $s_{kk'} = \mathbf{I}_k \cdot \mathbf{D} \cdot \mathbf{I}_{k'} / \sum \mathbf{I}_k \sum \mathbf{I}_{k'}$. Nous montrons en Fig. 3 les valeurs de ces différentes mesures, ainsi que la composition sémantique des communautés de citation, pour les classes sémantiques majoritaires. La distribution de I_i montre que les papiers gravitant dans le domaine du LUTI sont les plus interdisciplinaires dans les termes utilisés, ce qui pourrait être lié à leur caractère appliqué. Les autres disciplines sont dans des motifs similaires, à part la géographie et la planification des infrastructures qui présentent des distributions quasi-uniformes, témoignant de l'existence de références très spécialisées dans ces classes. Ce n'est pas nécessairement étonnant vu les sous-champs pointus exhibés (sciences politiques par exemples, et de même les études prospectives type coût-bénéfices sont très étriquées). Ce premier croisement des couches nous confirme les spécificités de chaque champ. Concernant les compositions sémantiques, la plupart agissent comme validation externe vu les classes majoritaires. Le champ le moins concerné par les problème socio-économiques est la planification des infrastructure, ce qui donnera du grain à moudre aux détracteurs de la technocratie. Les questions de changement climatique et durabilité sont relativement bien répartie. Enfin, les ouvrages géographiques concernent en majorité des problèmes de gouvernance. Les matrices de proximité confirment la conclusion de la sous-section précédente en terme de citation, les partages étant très faibles, les plus hautes valeurs étant jusqu'à un quart de la planification vers la géographie et des LUTI vers le TOD (mais pas l'inverse, les relations peuvent être à sens unique). Hors, les proximités sémantiques montrent par exemple que LUTI, TOD, Accessibility et Networks sont proches dans leur termes, ce qui est logique pour les trois premiers, et confirme pour le dernier que les physiciens se basent majoritairement sur les méthodes des ces champs liés au planing pour légitimer leur travaux. La géographie est totalement isolée, sa plus proche voisine étant la planification des infrastructures. Cette étude est très utile pour notre propos, puisqu'elle montre des domaines cloisonnés partageant des termes at donc a priori des problématiques et sujet commun. On ne se parle pas alors qu'on parle des languages pas si lointains, d'où la pertinence accrue de les faire parler d'une commune voie dans nos travaux : nos modèles devront mobiliser des éléments, ontologies et échelles de ces différents champs.

Nous concluons cette analyse par une approche plus robuste pour quantifier les proximités entre couches de l'hyperréseau. Il est aisé de construire une matrice de corrélation entre deux classifications, par les corrélations de leur colonnes. Nous définissons les probabilités $\mathbf{P}_C$ toutes égales à 1 pour la classification de citation. La matrice de correlation de celle-ci avec $\mathbf{P}$ s'étend de -0.17 à 0.54 et a une moyenne

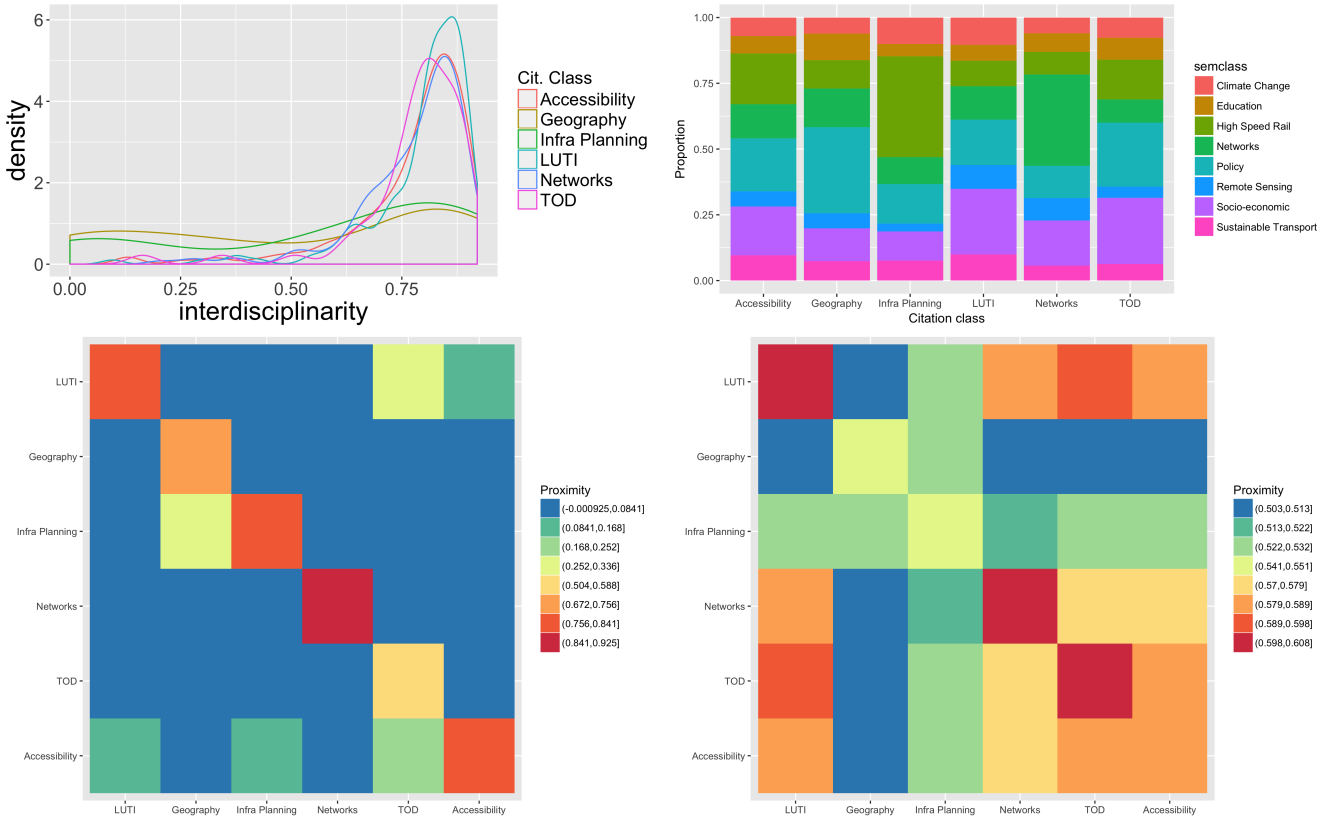


FIGURE 3 : **Motifs d'interdisciplinarité.** *(Haut Gauche)* Distribution des I_i par classes de citations ; *(Haut Droite)* Composition sémantiques des classes de citation ; *(Bas Gauche)* Matrice de proximité de citation $c_{kk'}$ entre classes de citations ; *(Bas Droite)* Matrice de proximité sémantique $s_{kk'}$ entre classes de citations.

de valeur absolue de 0.08, ce qui est significatif par rapport à des classifications aléatoire puisque un bootstrap à $b = 100$ répétitions avec les matrices mélangées donne un minimum à -0.08 ± 0.012, un maximum à 0.11 ± 0.02 et une moyenne absolue à 0.03 ± 0.002. Cela montre que les classifications sont complémentaires et que cette complémentarité est significative statistiquement par rapport à des classifications aléatoires. L'adéquation de la classification sémantique par rapport au réseau de citation peut également être quantifiée par la modularité multi-classes [Nicosia et al., 2009] (voir **??** pour une définition mathématique), qui traduit la probabilité qu'un lien soit dû à la classification étudiée, en prenant en compte l'appartenance simultanée à de multiples classes. Ainsi, la modularité multi-classes des probabilités sémantiques pour le réseau de citation est de 0.10, ce qui d'une part est significativement signe d'adéquation, un bootstrap toujours à $b = 100$ donnant une valeur de 0.073 ± 0.003, qui reste limitée vu la valeur maximale fixée par les probabilités de citations dans leur propre réseau qui donnent une valeur de 0.81, ce qui confirme d'autre part la complémentarité des classifications.

2.2.3 *Discussion*

Vers une modélisation des thèmes et une extraction automatique du contexte

Une direction possible pour renforcer cette analyse en épistémologie quantitative serait de travailler sur les textes complets des références contenant des efforts de modélisations des interactions entre réseaux et territoires, avec le but d'extraire automatiquement les thématiques des articles. Des méthodes plus adaptées pour les long texte que celle utilisée ici incluent par exemple l'Allocation Latente de Dirichlet [Blei, Ng et Jordan, 2003]. L'idée serait de procéder à une sorte de modélographie automatique, pour extraire des caractéristiques telle les ontologies, l'architecture ou la structure des modèles, les échelles ou même des valeurs typiques des paramètres. Il n'est pas clair dans quelle mesure la structure des modèles peut être extraite de leur description dans un article, et cela dépend sûrement de la discipline considérée. Par exemple dans champ relativement cadré comme la planification des transports, l'utilisation d'une ontologie pré-définie (dans le sens d'un dictionnaire) et d'une grammaire floue pourrait être efficace vu les conventions assez strictes dans la discipline. En géographie théorique et quantitative, au delà de la barrière du language, l'organisation de l'information est sûrement plus délicate à appréhender par de l'apprentissage non-supervisé à cause de la nature plus littéraire de la discipline : les synonymes et les figures de style sont généralement la norme pour l'écriture d'un bon niveau en sciences humaines, rendant plus floue une possible structure générique de la description des connaissances.

Réflexivité

La méthodologie que nous avons développé ici est particulièrement intéressante puisqu'elle offre des potentialités de réflexivité, c'est à dire qu'elle peut être utilisée pour étudier notre approche elle-même. Une de ses applications, hors de celle à la revue scientifique Cybergeo dans la perspective de Science Ouverte (voir Appendice B.6), sera à notre propre corpus de références, dans le but de révéler des possibles directions de recherche ou problématiques exotiques. Il est éventuellement possible de le faire de manière dynamique, grâce à l'historique de `git` qui permet de récupérer n'importe quelle version de la bibliographie à une date donnée sur les trois ans écoulés. Il s'agira aussi de comprendre nos motifs de production de connaissance afin de contribuer à 9.3. Le développement détaillé est fait en Appendice F.

⋆ ⋆

⋆

2.3 REVUE SYSTÉMATIQUE ET MODÉLOGRAPHIE

Tandis que les études menées précédemment proposaient de construire un horizon global de l'organisation des disciplines s'intéressant à notre question, nous proposons à présent une étude plus ciblée des caractéristiques de modèles existants. Nous proposons pour cela dans un premier temps une revue systématique, c'est à dire la construction d'un corpus plus précis répondant à certaines contraintes, suivie d'une meta-analyse, c'est à dire une tentative d'explication de certaines caractéristiques des modèles par des modèles statistiques.

2.3.1 *Revue systématique et Meta-analyse*

Les revues systématiques classiques ont majoritairement lieu dans des domaines où une recherche très ciblée, même par titre d'article, fournira un certain nombre d'études étudiant quasiment la même question : typiquement en évaluation thérapeutique, où des études standardisées d'une même molécule varient uniquement par taille des effectifs et modalités statistiques (groupe de contrôle, placebo, niveau d'aveugle). Dans ce cas la construction du corpus est d'une part aisée par l'existence de bases spécialisées permettant des recherches très ciblées, et d'autre part par la possibilité de procéder à des analyses statistiques supplémentaires pour croiser les différentes études (par exemple meta-analyse par réseau, voir [RUCKER, 2012]). Dans notre cas, l'exercice est bien plus aléatoire pour les raisons exposées dans les deux sections précédentes : les objets sont hybrides, les problématiques diverses, et les disciplines variées. Les différents points soulevés par la suite auront souvent autant de valeur thématique que de valeur méthodologique, suggérant des points cruciaux lors de la réalisation d'une telle revue systématique hybride.

Nous proposons une méthodologie hybride couplant les deux méthodologies développées précédemment avec une procédure plus classique de revue systématique. Nous souhaitons à la fois une représentativité de l'ensemble des disciplines que l'on a découvertes, mais aussi un bruit limité dans les références prises en compte pour la modélographie. Nous adoptons pour cela le protocole suivant :

1. Partant du corpus de citation isolé en 2.2.2, nous isolons un nombre de mots-clés pertinents, en sélectionnant les 5% de liens ayant le plus fort poids, puis parmi les noeuds correspondants ceux ayant un degré supérieur au quantile à 0.8 de leur classe sémantique respective. Le premier filtrage permet de se concentrer sur le "coeur" des disciplines observées, et le second de ne pas biaiser par la taille sans perdre la structure globale, les classes étant relativement équilibrées. Un examen manuel permet de supprimer les mots-clés clairement non-pertinents (télé-

détection, tourisme, réseaux sociaux, ...), ce qui conduit à un corpus de $K = 115$ mots-clés.

2. Pour chaque mots-clé, nous effectuons automatiquement une requête au catalogue (scholar) en y ajoutant `model*`, d'un nombre fixé $n = 20$ de références. L'ajout du terme est nécessaire pour obtenir des références pertinentes, après test sur des échantillons.

3. Le corpus potentiel composé des références obtenues, ainsi que des références composant le réseaux de citation, est revu manuellement (passage en revue des titres) pour assurer une pertinence au regard de l'état de l'art de 2.1, fournissant le corpus préliminaire de taille $N_p = 297$.

4. Ce corpus est alors inspecté pour les résumés et textes complets si nécessaire. On sélectionne les articles mettant en place une démarche de modélisation, hors modèles conceptuels. Les références sont classifiées et caractérisées selon des critères décrits ci-dessous. On obtient alors un corpus final de taille $N_f = 145$, sur lequel des analyses quantitatives sont possibles.

La méthode est résumée en Fig. 4, avec les valeurs des paramètres et la taille des corpus successifs. Cet exercice permet tout d'abord un certain nombre de points méthodologiques, dont la connaissance pourra être un atout pour mener des revues systématiques hybrides similaires :

- Les biais de catalogue semblent inévitables. Nous reposons sur l'hypothèse que l'utilisation de Scholar permet un échantillonnage uniforme au regard des erreurs ou biais de catalogage. Le développement futur d'outils ouverts de catalogage et de cartographie, permettant un effort contributif pour une connaissance plus précise de domaines étendus et de leurs interfaces, sera un enjeu crucial de la fiabilité de ce genre de méthodes (voir B.6).

- La disponibilité des textes complets est particulièrement un problème pour une revue si large, vu la multiplicité des éditeurs. L'existence de moyens d'émancipation de la science ouverte comme Sci-hub permet d'effectivement accéder à l'ensemble des textes. En écho au débat sur le bras de fer récent avec les éditeurs concernant l'exclusivité de la fouille de textes complets, il parait de plus en plus évident qu'une science ouverte réflexive est totalement antagoniste au modèle actuel de l'édition. Nous espérons également une évolution rapide des pratiques sur ce point.

- Les revues, et en fait les éditeurs, semblent influencer différemment les référencements, augmentant potentiellement le biais de requête. La littérature grise ainsi que les pre-prints sont pris en compte différemment selon les champs.

- Le passage en revue manuel des grand corpus permet de pas louper des "poids lourds" qui auraient pu être omis en amont [Lissack, 2013]. La question de la mesure dans laquelle on peut s'attendre d'être au courant de la manière la plus exhaustive des découvertes récentes liées au sujet étudié évolue très probablement vu l'augmentation de la quantité totale de littérature produite et la fragmentation des domaines pour certains toujours plus pointus [Bastian, Glasziou et Chalmers, 2010]. Rejoignant les points précédents, on peut supposer que des outils d'aide à l'analyse systématique permettront de garder cet objectif raisonnable.

- Les résultats de la revue automatique sont sensiblement différents des domaines dessinés dans la revue classique : certaines associations conceptuelles, notamment l'inclusion des modèles de croissance de réseaux, ne sont pas naturelles et existent peu dans le paysage scientifique comme nous l'avons montré précédemment.

D'autre part, l'opération de construction du corpus permet déjà en elle-même de tirer des observations thématiques intéressantes en elles-mêmes :

- Les articles sélectionnés supposent une clarification de ce qui est entendu par "modèle". Nous donnons en 9.3 une définition très large s'appliquant à l'ensemble des perspectives scientifiques. Notre selection ici ne retient pas les modèles conceptuels par exemple, notre critère de choix étant que le modèle doit inclure un aspect numérique ou de simulation.

- Un certain nombre de références consistent en des revues, ce qui revient à un groupe de modèles ayant des caractéristiques similaires. On pourrait compliquer la méthode en retranscrivant chaque revue ou meta-analyse, ou en pondérant par le nombre d'article correspondant les enregistrements des caractéristiques correspondants. Nous faisons le choix d'ignorer ces revues, ce qui reste cohérent de manière thématique en restant dans l'hypothèse d'échantillonnage uniforme.

- Une première clarification du cadre thématique est opérée, puisque nous ne sélectionnons pas les études liées uniquement au traffic et à la mobilité (ce choix étant aussi lié aux résultats obtenus en 5.1), à l'urban design pur, au modèles de flux piétons, au fret, à l'écologie, aux aspects techniques du transport, pour donner quelques exemples, même si ces sujets peuvent dans une vue extrême être considérés comme liés aux interactions entre réseaux et territoires.

- De la même façon, des domaines annexes comme le tourisme, les aspects sociaux de l'accès aux transports, l'anthropologie, n'ont pas été pris en compte.
- On observe une forte fréquence des études liées au Trains à Grande Vitesse (HSR), rappelant la non-dissociabilité des aspects politiques de la planification et des directions de recherche en transports, surtout en France où les Corpsards des Ponts ou des Mines ont une main mise relative sur les deux aspects simultanément.

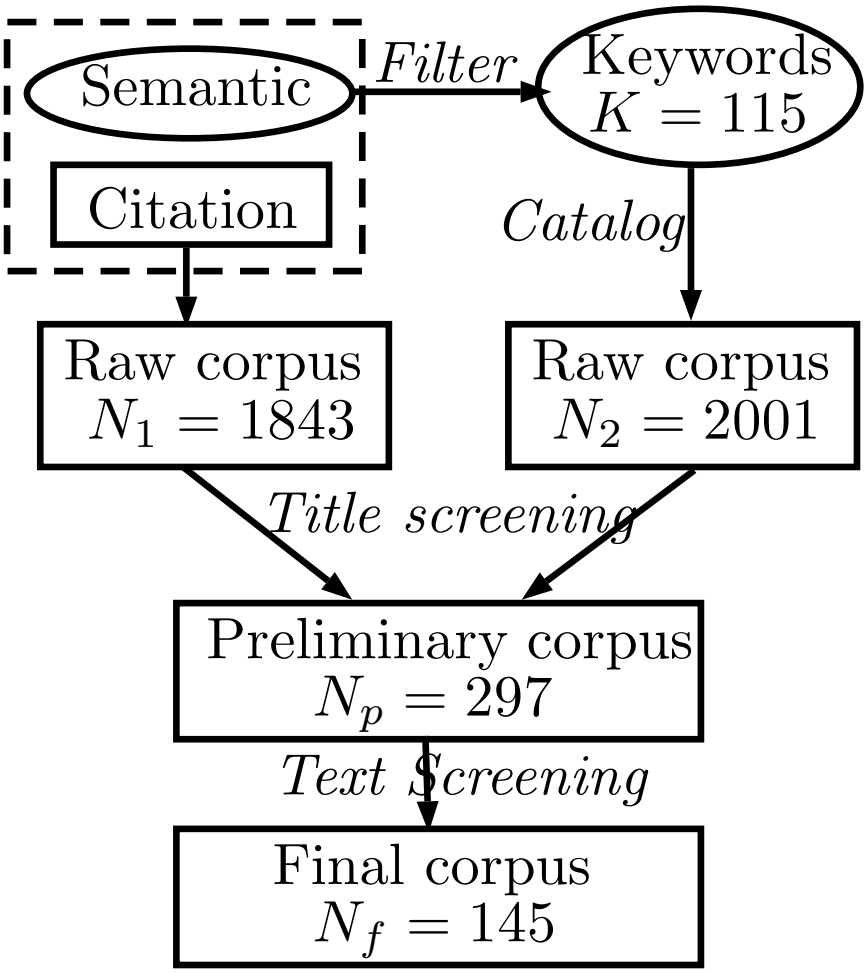


FIGURE 4 : **Méthodologie de la revue systématique.**

2.3.2 *Modélographie*

Nous passons à présent à une analyse mixte basée sur ce corpus, inspirée par les résultats des sections précédentes précédents notamment pour la classification. Elle a pour but d'extraire et de décomposer précisément les ontologies, échelles et processus, puis d'étudier des liens

possibles entre ces caractéristiques des modèles et le contexte dans lequel ils ont été introduits. Il s'agit ainsi de la meta-analyse en quelque sorte, que nous désignerons ici par modélographie. Pour ne pas froisser les puristes, il ne s'agit en effet pas d'une meta-analyse à proprement parler car nous ne combinons pas des analyses proches pour extrapoler des résultats potentiels d'échantillons plus grand. Notre démarche est proche de celle de COTTINEAU dans [COTTINEAU, 2016] qui rassemble les références ayant étudié quantitativement la loi de Zipf pour les villes, puis lie les caractéristiques des études aux méthodes utilisées et hypothèses formulées.

La première partie consiste en l'extraction des caractéristiques des modèles. Automatiser ce travail constituerait un projet de recherche en lui-même, comme nous développons en discussion ci-dessous, mais nous sommes convaincus de la pertinence d'affiner de telles techniques (voir 9.3.3) dans le cadre d'un développement de disciplines intégrées. Le temps étant autant l'ennemi que l'allié de la recherche, nous nous concentrons ici sur une extraction manuelle qui se voudra plus fine qu'une tentative peu convaincante de fouille de données. Nous extrayons des modèles les caractéristiques suivantes :

- Quelle est la force du couplage entre les ontologies territoriales et celles du réseau, autrement dit s'agit-il d'un modèle de coévolution. Nous classerons pour cela en catégories suivant la représentation de la figure 5 : `{territory ; network ; weak ; coevolution}`, qui résulte de l'analyse de la littérature en 2.1.
- Echelle de temps maximale.
- Echelle d'espace maximale.
- Hypothèses d'équilibre.
- Domaine "a priori", déterminé par l'origine des auteurs et domaine de la revue.
- Méthodologie utilisée (modèles statistiques, système d'équations, multi-agent, automate cellulaire, recherche opérationnelle, simulation etc.).
- Cas d'étude (ville, métropole, région ou pays) s'il y a lieu.

Nous collectons également de manière indicative, mais sans objectif d'objectivité ni d'exhaustivité, le "sujet" de l'étude (c'est à dire la question thématique dominante) ainsi que les "processus" inclus dans le modèle. Une extraction exacte des processus reste hypothétique, d'une part conditionnée à une définition rigoureuse et prenant en compte différents niveaux d'abstraction, de complexité, ou d'échelle, d'autre part dépendant de moyens techniques hors de portée de cette étude modeste. Nous commenterons ceux-ci de manière indicative sans les inclure dans les études systématiques.

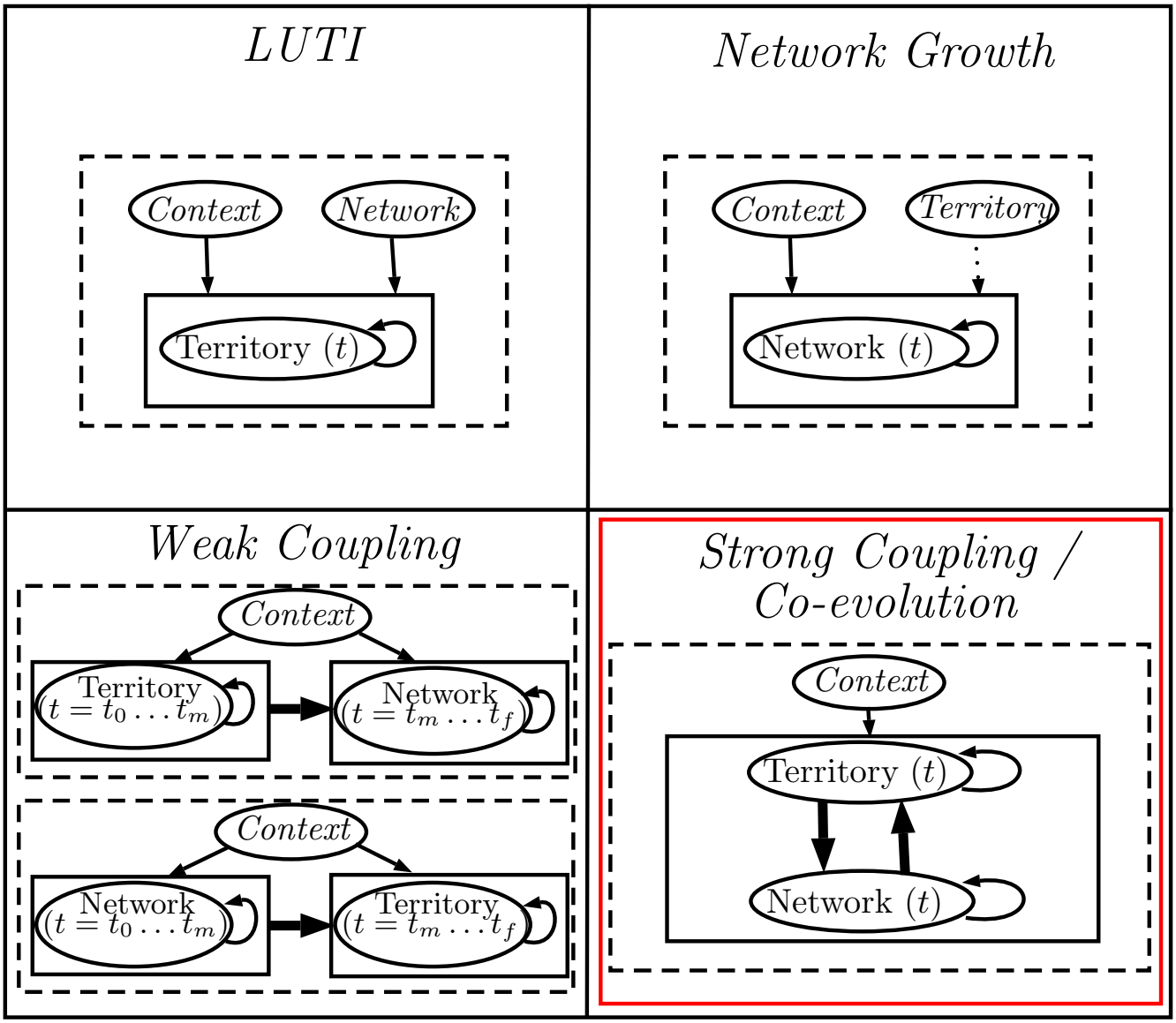


FIGURE 5 : **Représentation schématique de la distinction entre différents types de modèles couplant territoires et réseaux.** Les ontologies sont représentés par des ovales, les sous-modèles par les boîtes pleines, les modèles par les boîtes pointillées, les couplages par les flèches. Nous surlignons en rouge l'approche qui sera l'objectif final de notre travail.

Nous confondons également échelle, portée et dans un sens résolution pour ne pas rendre plus confus l'extraction. Même s'il serait pertinent de différencier lorsque un élément n'a pas lieu d'être pour un modèle (NA) de lorsque celui-ci est mal défini par son auteur, cette tâche apparaît sujette à subjectivité et nous fusionnons les deux modalités. Nous ajoutons aux caractéristiques ci-dessus les variables suivantes :

- Domaine de citation (le cas échéant, c'est à dire pour les références initialement présentes dans le réseau de citation, i.e. 55% des références)
- Domaine sémantique, défini par le domaine pour lequel le document a la plus grande probabilité
- Indice d'interdisciplinarité

Les domaines sémantiques et la mesure d'interdisciplinarité ont été recalculés pour ce corpus par collecte des mots-clés, puis extraction selon la méthode décrite en 2.2, avec $K_W = 1000$, $\theta_w = 15$ et $k_{max} = 500$. On obtient des communautés plus ciblées et plutôt représentatives de la thématique et des méthodes : Transit-oriented development (`tod`), Hedonic models (`hedonic`), Planification des infrastructures (`infra planning`), High-speed rail (`hsr`) , Réseaux (`networks`), Réseaux complexes (`complex networks`), Bus rapid transit (`brt`).

Un "bon choix" de caractéristiques pour classer les modèles est un peu le problème du choix des *features* en apprentissage statistique : si on est en supervisé, c'est à dire qu'on veut obtenir une bonne prédiction de classe fixée a priori (ou une bonne modularité de la classification obtenue par rapport à la classification fixée), on pourra sélectionner les caractéristiques optimisant cette prédiction. On discriminera ainsi les modèles que l'on connait et que l'on juge différents. Si l'on veut extraire une structure endogène sans a priori (classification non supervisée), la question est différente. Nous testerons pour cela en second temps une technique de regression qui permet d'éviter l'overfitting et faire de la selection de caractéristiques (Forêts aléatoires).

PROCESSUS ET CAS D'ÉTUDE Concernant l'existence d'un cas d'étude et sa localisation, 26% des études n'en présentent pas, correspondant à un modèle abstrait ou modèle jouet (la quasi totalité des études en physique tombant dans ce cas). Ensuite, elles sont réparties à travers le monde, avec toutefois une surreprésentation des Pays-bas avec 6.9%. Les processus inclus sont trop variés (en fait autant que les ontologies des disciplines concernées) pour faire l'objet d'une typologie, mais on notera la domination de la notion d'accessibilité (65% des études), puis des processus très variés allant de processus de marché immobilier pour les études hédoniques, aux relocalisations d'actifs et d'emplois pour les luti, ou aux investissements d'infrastructure de réseau. On observe des processus abstraits géométriques de croissance de réseau, correspondant aux travaux des physiciens. La maintenance du réseau apparait dans une étude, ainsi que l'histoire politique. Les processus abstraits d'agglomération et dispersion sont aussi le coeur de quelques études. Les interactions entre villes sont minoritaire, les approches de type système de villes étant noyées dans les études d'accessibilité. Les questions de gouvernance et de régulation ressortent aussi, plutôt dans le cas de planification d'infrastructure et de modèle

d'évaluation de démarches TOD, mais sont aussi minoritaires. On retiendra que chaque domaine puis chaque étude introduit ses propres processus quasi-spécifiques à chaque cas.

CARACTÉRISTIQUES DU CORPUS Les domaines "a priori" (i.e. jugés, ou plutôt préjugés sur la revue ou l'appartenance des auteurs), sont relativement équilibrés pour les disciplines majoritaires déjà identifiées : 17.9% Transportation, 20.0% Planning, 30.3% Economics, 19.3% Geography, 8.3% physics, le reste minoritaire se répartissant entre environnement, informatique, ingénierie et biologie. Concernant les poids des domaines sémantiques significatifs, le TOD domine avec 27.6% des documents, suivi par les réseaux (20.7%), les modèles hédoniques (11.0%), la planification des infrastructures (5.5%) et le HSR (2.8%). Les contingences montrent que le Planning ne fait quasiment que du TOD, la physique uniquement des réseaux, la géographie se répartit équitablement entre réseaux et TOD (le second correspondant aux articles typés "aménagement", qui ont été classés en géographie car dans des revues de géographie) ainsi qu'une plus faible part en HSR, enfin l'économie est la plus variée entre hédonique, planning, réseaux et TOD. Cette interdisciplinarité n'apparait cependant que pour les classes extraites pour la probabilité majoritaire, puisque les indices d'interdisciplinarité moyens par discipline ont des valeurs équivalentes (de 0.62 à 0.65), hormis la physique significativement plus basse à 0.56 ce qui confirme son statut de "nouveau venu" ayant une profondeur thématique plus faible.

Il est intéressant pour notre question de répondre à la question "qui fait quoi ?", c'est à dire quelles types de modèles sont mobilisés par les différentes disciplines. Nous donnons en Table **??** la table de contingence du type de modèle en fonction des disciplines a priori, de la classe de citation et de la classe sémantique. On constate les approches fortement couplées, les plus proches de ce qu'on considère comme des modèles de co-évolution, sont majoritairement contenues dans le vocabulaire des réseaux, ce qui est confirmé par leur positionnement en terme de citation, mais que les disciplines concernées sont variées. La majorité des études s'intéresse au territoire uniquement, le déséquilibre le plus fort étant pour les études sémantiquement liées au TOD et à l'hédonique. La physique est encore limitée en s'interessant exclusivement aux réseaux.

Pour répondre ensuite à la question du comment, on peut regarder les échelles de temps et d'espace typiques des modèles. La planification et les transports se concentrent à des petites échelles spatiales, métropolitain ou local, l'économie également avec une forte représentation du local via les études hédoniques, et une étendue un peu plus grande avec l'existence d'études au niveau régional et quelques une du pays (études de panel généralement). Encore une fois, la physique se retrouve limitée avec l'ensemble de ses contributions à une échelle

TABLE 3 : **Type de modèles étudiés selon les différentes classifications.** Tables de contingence de la variable discrete donnant le type de modèle (réseau, territoire ou couplage fort), pour la classification a priori, la classification sémantique et la classification de citation.

Discipline	economics	geography	physics	planning	transportation
network	5	3	12	1	4
strong	4	3	0	0	2
territory	35	22	0	28	20

Semantic	hedonic	hsr	infra planning	networks	tod
network	1	0	0	14	2
strong	0	0	0	5	1
territory	15	4	8	11	37

Citation	Accessibility	Geography	Infra Planning	LUTI	Networks	TOD
network	0	0	0	0	24	0
strong	0	0	0	2	5	0
territory	13	1	6	18	2	3

fixe, métropolitaine (pas forcément claire ni bien spécifiée dans les articles d'ailleurs puisqu'il s'agit de modèles jouets dont les contours thématiques peuvent être très flous). La géographie est relativement bien équilibrée, de l'échelle métropolitaine à l'échelle continentale. Le schéma pour les échelles de temps est globalement similaire. Les méthodes utilisées sont fortement corrélées à la discipline : un test du χ^2 donne une statistique de 169, très significatif avec $p = 0.04$. De même, l'échelle d'espace l'est mais de manière moindre ($\chi^2 = 50, p = 0.08$).

RÉGRESSIONS CLASSIQUES Nous étudions l'influence de divers facteurs sur les caractéristiques des modèles par des régressions linéaires simples. Dans une démarche de multi-modélisation, nous testons l'ensemble des modèles possible pour expliquer la variable à partir des autres, et sélectionnons le meilleur en terme de Critère d'Information d'Akaike. Les résultats complets des régressions sont donnés en Appendice A.3. L'échelle temporelle et d'espace sont les mieux expliquées par les modèles prenant en compte l'ensemble des autres variables. Pour l'échelle de temps, les variables les plus significative sont le fait d'utiliser des méthodes de simulation et le fait d'être en physique, qui tous deux influent négativement. L'échelle spatiale et le fait d'être en planification influent positivement. Au contraire pour l'échelle d'espace, le fait d'être en planning influence négativement alors que le domaine sémantique du TOD est positif, ce qui veut dire

que les journaux de planning privilégient des études localisées alors des problématiques proches ont tendance à étendre l'aire d'étude. Le niveau d'interdisciplinarité est le mieux expliqué par une unique variable, l'année, qui l'influence de manière négative, ce qui confirme l'augmentation des spécialisations scientifiques dans le temps.

RÉGRESSIONS PAR FORÊTS ALÉATOIRES Nous concluons cette étude par des régressions et classification par Forêts Aléatoires, qui sont une méthode très flexible permettant de dégager une structure d'un jeu de données [LIAW et WIENER, 2002]. Pour compléter les analyses précédentes, nous proposons de l'utiliser pour déterminer les importances relatives des variables pour différents aspects. Nous utilisons à chaque fois des forêts de taille 100000, une taille de noeud de 1 et un nombre de variable échantillonnée en $\sqrt{p}$ pour la classification et $p/3$ pour la régression lorsque p est le nombre total de variables. Pour classifier le type de modèle, nous comparons les effets de la discipline, de la classe sémantique et de la classe de citation. Cette dernière est la plus importante avec une mesure relative de 45%, tandis que la discipline compte pour 31% et le sémantique pour 23%. Ainsi, le cloisonnement disciplinaire se retrouve, tandis que le sémantique et donc en partie les ontologies, est le plus ouvert. Cela nous encourage dans notre démarche de sortir de ce cloisonnement. Lorsqu'on applique une regression de forêt sur l'interdisciplinarité, toujours avec ces trois variables, on constate qu'elles expliquent 7.6% de la variance totale, ce qui est relativement faible, témoignant d'une disparité de sémantique sur l'ensemble du corpus indépendamment des différentes classifications. Dans ce cas, la variable la plus importante est la discipline (39%) suivie par le sémantique (31%) et la citation (29%), ce qui confirme que le journal visé conditionne fortement le comportement de langage employé. Cela nous alerte sur le danger de perte de richesse sémantique lorsqu'on s'adresse à un public particulier. Ainsi, nous avons pu dégager certaines structures et régularités des modèles nous concernant, qui seront riches d'enseignements lors de la construction de nos modèles.

2.3.3 *Discussion*

DÉVELOPPEMENTS Un développement possible pourrait consister en la mise en place d'une approche automatique à cette meta-analyse, du point de vue de la modélisation modulaire, combiné avec une classification du but et de l'échelle. La modélisation modulaire consiste en l'intégration de processus hétérogènes et d'implémentation de ces processus dans le but d'extraire les mécanismes donnant la meilleure proximité à des faits stylisés empiriques ou à des données [COTTINEAU, CHAPRON et REUILLON, 2015]. L'idée serait de pouvoir extraire automatiquement la structure modulaire des modèles existants, à par-

tir des textes complets comme proposé en 2.2, afin de classifier ces briques de manière endogène et identifier des couplages potentiels pour des nouveaux modèles.

LEÇONS POUR LA MODÉLISATION Nous pouvons résumer les points principaux issus de cette méta-analyse qui joueront sur notre attitude et nos choix de modélisation. Tout d'abord, la présence interdisciplinaire des approches effectuant un couplage fort confirme notre besoin de faire des ponts et de coupler les approches, et confirme également rétrospectivement les conclusions de 2.2 sur les conséquences du cloisonnement des disciplines en terme de modèles formulés. Ensuite, l'importance du vocabulaire des réseaux dans une grande partie des modèles nous poussera à confirmer cet ancrage. La spécificité des approches TOD et d'accessibilité, assez proches des modèles LUTI, seront secondaires pour nous. La portée restreinte des travaux issus de la physique, confirmée par la majorité des critères étudiés, nous pousse à nous méfier de ces travaux et de l'absence de sens thématique aux modèles. La richesse des échelles temporelles et spatiale couvertes par les modèles géographiques et économiques nous confirme l'importance de varier celles-ci dans nos modèles, idéalement de parvenir à des modèles multi-échelles. Enfin, les importances relatives des variables de classification sur le type de modèle vont également dans le sens de ponts interdisciplinaires pour croiser les ontologies.

⋆ ⋆

⋆

CONCLUSION DU CHAPITRE

La réflexivité semble dans notre cas être nécessaire pour une appréhension claire des enjeux thématiques, méthodologiques et plus généralement scientifiques liés au processus que nous cherchons à modéliser : ceux-ci étant multi-scalaires, hybrides et hétérogènes, les angles d'approches et questionnements possibles sont nécessairement extrêmement variés, complémentaires et riche. Il pourrait s'agir d'une caractéristique fondamentale des systèmes socio-techniques, que PUMAIN formule dans [PUMAIN, 2005] comme "une nouvelle mesure de complexité", qui serait liée aux nombre de point de vue nécessaires pour appréhender un système à un niveau donné d'exhaustivité. Cette idée rejoint la position de *perspectivisme appliqué* que la section 9.2 formalise et qui est implicitement présente dans l'investigation des relations entre Economie et Géographie développée en C.1. Ainsi, la modélisation des interactions entre réseaux et territoires peut être reliées à un ensemble très large de disciplines et d'approches revues en section 2.1. Afin de mieux comprendre le paysage scientifique environnant, et quantifier les rôles ou poids relatifs de chacune, nous avons procédé à une série d'analyse en épistémologie quantitative en 2.2. Une première analyse préliminaire basée sur une revue systématique algorithmique suggère un certain cloisonnement des domaines. Cette conclusion est confirmée par l'analyse d'hyperréseau couplant réseau de citation et réseau sémantique, qui permet également de dessiner plus finement les contours disciplinaires, à la fois sur leur relations directes (citations) mais aussi leur proximité scientifique pour les termes et méthodes utilisées. On peut alors utiliser le corpus constitué et cette connaissance des domaines pour une revue systématique semi-automatique et une meta-analyse en 2.3, qui permet de constituer un corpus de travaux traitant directement du sujet, qui est ensuite inspecté intégralement, permettant de lier caractéristique des modèles au différents domaines. On a alors à ce stade une idée assez précise de ce qui ce fait, pourquoi et comment. L'enjeu reste de déterminer les pertinences relatives de certaines approches ou ontologies, ce qui sera le but des trois chapitres de la deuxième partie. Nous concluons d'abord cette première partie par un chapitre de discussion 3, éclairant des points nécessaires à clarifier avant une entrée dans le vif du sujet.

⋆ ⋆

⋆

3

POSITIONNEMENTS

Toute activité de recherche serait, selon certains observateurs, nécessairement politisée, de par pour commencer le choix de ses objets. Ainsi, Ripoll alerte contre l'illusion d'une recherche objective et les dangers de la technocratie [Ripoll, 2017]. Nous ne rentrerons pas dans ces débats bien trop vastes pour être traités même en un chapitre, puisqu'il rejoignent des thèmes de sciences politiques, d'éthique, de philosophie, liés par exemple à la gouvernance scientifique, à l'insertion de la science dans la société, à la responsabilité scientifique. Il est clair que même des sujets a priori intrinsèquement objectifs, comme la physique des particules et des hautes énergies, ont des implications regardant d'une part les choix de leur financements et les externalités associées (par exemple, l'existence du CERN a largement contribué au développement du calcul distribué), mais d'autre part aussi les applications potentielles des découvertes qui peuvent avoir des répercussions sociales considérables. En biologie, l'éthique est au coeur des principes fondateurs des disciplines, comme en témoignent les débats soulevés par l'émergence de la biologie synthétique [Gutmann, 2011]. Les tenants d'approche prudentes dans celle-ci se recoupent avec la biologie intégrative, or les Sciences Intégratives défendues par Paul Bourgine, mises en oeuvre par l'intermédiaire du campus digital Unesco CS-DC[1], ont typiquement la responsabilité sociale et l'implication citoyenne au coeur de leur cercle vertueux. En sciences humaines, comme les recherches interagissent avec les objets étudiés (en quelque sorte l'idée des *interactive kind* de Hacking [Hacking, 1999]), les implications politiques et sociales de la recherche sont bien évidemment indiscutables. Là où il y aurait matière à discussion, et nous y reviendrons en ouverture 9.3.3 car il s'agira d'une des questions ouvertes posées par notre recherche et sa démarche dans leur ensemble, serait sur la compatibilité des méthodes systématiques et *evidence-based* avec les sciences sociales, autrement dit dans quelle mesure peut-on s'extraire de certains dogmatismes encore plus marqués lors de l'usage de théorie politiques[2]. Nous resterons ici à un niveau épistémologique, c'est à dire à des réflexions sur la nature et le contenu des connaissances scientifiques au sens large, c'est à dire coconstruites et validées au sein d'une communauté imposant certains critères de scientificité, bien sûr évolutifs puisque nous nous posi-

1 https ://www.cs-dc.org/

2 Monod montre par exemple les désastres liés aux "niaiseries épistémologiques" découlant de l'application littérale de la dialectique matérialiste marxiste à l'épistémologie du vivant.

tionnerons pour la systématisation de certains. Mais donc, même en restant à ce niveau, des prises de positions sont nécessaires, celles-ci pouvant être épistémologiques, méthodologiques, thématiques. Ces dernières ont déjà été ébauchée dans les deux chapitres précédents par les choix des objets d'étude, des problématiques, et seront renforcées à mesure de la progression pour finalement être synthétisées en Chapitre 9. Nous proposons ici un exercice relativement original mais que nous jugeons nécessaire pour une lecture plus fluide de la suite, qui consiste en le développement précis de certains positionnements qui ont une influence particulière dans notre démarche de recherche. Par exemple, le travail en données quasi-intégralement ouverte et en architecture modulaire résulte de notre exigence de reproductibilité. L'utilisation des modèles et la manière de les explorer de notre vision du calcul intensif. Dans une première section (3.1), nous développons des exemples pour illustrer le besoin et la difficulté de reproductibilité, ainsi que les liens avec des nouveaux outils pouvant la favoriser mais aussi la mettre en danger. Dans une deuxième section (3.2), nous argumentons sous forme d'essai pour un usage raisonné des données massives et du calcul intensif, et illustrons notre positionnement par rapport à l'exploration des modèles par une étude de cas méthodologique pour l'exploration de la sensibilité des modèles aux conditions initiales. Enfin, la dernière section (3.3) explicite modestement des positions épistémologiques, notamment concernant le courant dans lequel nous nous plaçons, la complexité des objets en sciences sociales, et la nature de la complexité de manière générale. Le lecteur très familier avec les commandements de Banos [Banos, 2013] pourra éventuellement sauter les deux premières sections à part s'il est intéressé par des illustrations pratiques originales, notre positionnement étant très similaire et ne divergeant que sur des subtilités mineures pour les sujets évoqués dans ces sections.

⋆ ⋆

⋆

Ce chapitre est composé de divers travaux. La première section est inédite. La deuxième section rend compte pour sa première partie du contenu théorique de [Raimbault, 2016b], et pour sa deuxième partie des idées présentées dans [Cottineau et al., 2017]. La troisième section reprend dans sa première partie les bases épistémologiques de [Raimbault, 2017g] approfondies par [Raimbault, 2017a], est inédite pour sa deuxième partie et rend compte de [Raimbault, 2017c] pour sa dernière partie.

3.1 REPRODUCIBILITÉ

La force de la Science vient de la nature cumulative et collective de la recherche, puisque les progrès sont faits lorsque, comme NEWTON l'a bien posé, on "se tient sur les épaules de géants", au sens que l'entreprise scientifique à un temps donné repose sur l'ensemble du travail précédent et qu'aucune avancée ne serait possible sans construire dessus. Cela inclut le développement de nouvelles théories, mais aussi l'extension, le test et la falsification de précédentes : l'avancée dans la construction de la tour signifie aussi la déconstruction de certaines briques obsolètes. Cet aspect de validation par les pairs et de remise en question constante est aussi ce qui légitime la Science pour une connaissance plus robuste et un progrès sociétal basés sur une connaissance d'un univers objectif, par rapport aux systèmes dogmatiques qu'ils soient politiques ou religieux [BAIS, 2010].

La reproductibilité semble être de plus en plus pratiquée de manière effective [STODDEN, 2010] et les moyens techniques pour l'achever sont toujours plus développés (comme par exemple les outils pour déposer les données ouvertes, ou pour être transparent dans le processus de recherche comme `git` [RAM, 2013], ou pour intégrer la création de document et l'analyse de données comme `knitr` [XIE, 2013]), au moins dans le champ de la modélisation et de la simulation. Cependant le diable est bien dans les détails et des obstacles jugés dans un premier temps comme mineurs peuvent rapidement devenir un fardeau pour reproduire et utiliser des résultats obtenus dans des recherches précédentes. Nous décrivons deux études de cas où les modèles de simulation sont en apparence hautement reproductibles mais se révèlent vite des puzzles pour lesquels l'équilibre de temps de recherche passe rapidement sous zéro, au sens où essayer d'exploiter leur résultats coûtera plus en temps que de développer entièrement des modèles similaires.

C : [CRICK, HALL et ISHTIAQ, 2015]

3.1.1 *Explicitation, documentation et implémentation des modèles*

Sur le Besoin d'expliciter le modèle

Un mythe à la vie dure (auquel nous essayons en fait nous-même d'échapper) est que fournir le code source complet et les données seront une condition suffisante pour la reproductibilité, puisque la reproductibilité computationnelle complète implique un environnement similaire ce qui devient vite ardu à produire comme le montre [HATTON et WARR, 2016]. Pour résoudre ce problème, [HUNG et al., 2016] propose l'utilisation de conteneurs Dockers qui permet de reproduire même le comportement de logiciels avec interface graphique indepedemment de l'environnement. C'est d'ailleurs une des direction courantes de développement d'OpenMole, pour simplifier le packa-

ging des bibliothèques et des modèles en binaire (cf. R. REUILLON dans [RAIMBAULT, 2017d]). Dans tous les cas, le reproductibilité a des dimensions supplémentaires, il ne s'agit pas de l'objectif unique qui serait est de produire exactement les mêmes graphes et analyses statistiques, en supposant que le code fournit est celui qui a été effectivement utilisé pour produire les résultats donnés. Tout d'abord, doivent être autant que possible indépendants de l'implémentation (c'est à dire du langage, des bibliothèques, des choix de structures de données et de type de programmation) pour des motifs clairs de robustesse. Ensuite, en relation avec le point précédent, un des buts de la reproductibilité est la réutilisation des méthodes ou résultats comme base ou modules pour une recherche future (ce qui comprend une implémentation dans un autre langage ou une adaptation de la méthode), au sens que la reproductibilité n'est pas la possibilité stricte de répliquer car elle doit être adaptable [DRUMMOND, 2009].

Notre premier cas d'étude suit exactement ce schéma, puisqu'il a sans aucun doute été conçu pour être partagé avec la communauté et utilisé, s'agissant d'un modèle de simulation fournit avec la plateforme de modélisation agent NetLogo [WILENSKY, 1999]. Le modèle est également disponible en ligne [DE LEON, FELSEN et WILENSKY, 2007] et est présenté comme un outil pour simuler les dynamiques socio-économiques des résidents à bas revenus d'une ville au sein d'un environnement urbain synthétique, généré pour ressembler en terme de faits stylisés à la ville réelle de Tijuana, Mexico. Globalement, le modèle fonctionne de la façon suivante : (i) à partir de centre urbains, une distribution d'usage du sol est générée par modélisation procédurale similaire à [LECHNER et al., 2006], c'est à dire des routes sont générées de proche en proche selon des règles géométriques et de hiérarchie locales, et un usage du sol ainsi qu'une valeur est attribué en fonction des caractéristique du patch (distance au centre, à la route) ; (ii) dans cet environnement urbain sont simulées des dynamiques résidentielles de migrants, qui cherchent à optimiser une fonction d'utilité dépendant du coût de la vie et de la configuration des autres migrants. A part fournir le code source, le modèle n'est que peu documenté dans la littérature ou dans les commentaires et la description de l'implémentation. Les commentaires qui suivent sont basés sur l'étude de la partie du modèle simulant la morphogenèse urbaine (setup pour la composante "dynamiques résidentielles") comme il s'agit de notre contexte global d'étude. Dans le cadre de cette étude, le code source a été modifié et commenté, dont la dernière version est disponible sur le dépôt du projet[3].

FORMALISATION RIGOUREUSE Une partie évidente de la construction d'un modèle est sa formalisation rigoureuse dans un cadre formel distinct du code source. Il n'y a bien sûr aucun langage universel

3 at `https://github.com/JusteRaimbault/CityNetwork/tree/master/Models/Reproduction/UrbanSuite`

pour le formuler [Banos, 2013], et de nombreuses possibilités sont offertes par de nombreux champs (e.g. UML, DEVS, formulation mathématique pure), mais l'étape de formalisation précise, qui suit généralement une description plus intuitive donnant les idées et processus dominants ("rationelle"), ne peut pas être sautée. On pourrait se dire que le code source y est équivalent, mais ce n'est pas exactement vrai car on pourrait alors ne plus distinguer certains choix d'implémentation de la structure du modèle. Aucun article ni documentation n'accompagne le modèle ici, au delà de la documentation embarquée NetLogo, qui ne décrit que de manière thématique en langage naturel les idées derrière chaque étape sans plus développer et fournit de l'information sur le rôle des différents éléments de l'interface. Comme ces éléments manquent ici, le modèle n'est guère utilisable tel quel. On pourrait nous objecter ici que la partie que nous étudions est une procédure d'initialisation et non le coeur du modèle : nous maintenons que l'ensemble des procédures doit être également documenté et implémenté avec un soin équivalent, ou pointer vers une référence extérieure dans le cas d'utilisation d'un modèle tiers, comme nous le faisons d'ailleurs pour le couplage effectué en 3.2.

Une telle formulation est essentielle pour que le modèle soit compris, reproduit et adapté ; mais elle évite également des biais d'implémentation comme

- Des éléments architecturaux dangereux : dans le modèle, le contexte du monde est une sphère, ce qui n'est pas raisonnable pour un modèle à l'échelle d'une ville. Les agents peuvent "sauter" dans la représentation euclidienne, ce qui n'est pas acceptable pour une projection en deux dimensions du monde réel. Pour éviter cela, de nombreux tests et fonctions subtils sont utilisés, incluant des pratiques déconseillées (e.g. mort d'agents basée sur leur position pour les empêcher de sauter).

- Manque de cohérence interne : par exemple la variable de patch `land-value` utilisée pour représenter différentes quantités géographiques à différentes étapes du modèle (morphogenèse et dynamiques résidentielles), ce qui devient une incohérence interne quand les deux étapes sont couplées lorsque l'option permettant de faire croître la ville est activée.

- Erreur de code : dans un langage non typé comme NetLogo, le mélange des types peut conduire à des erreurs inattendues à l'execution, ou même des *bugs* non détectables directement et alors plus dangereux. C'est le cas de la variable de patch `transport` dans le modèle (même si aucune erreur ne survient dans la majorité des configurations depuis l'interface, ce qui est plus dangereux comme le développeur pense que l'implémentation est sûre). De tels problèmes devraient être évités si

l'implémentation est faite à partir d'une description exacte du modèle.

IMPLÉMENTATION TRANSPARENTE Une implémentation totalement transparente doit être attendue, incluant une certaine ergonomie dans l'architecture et le code, mais aussi dans l'interface et la description du comportement attendu du modèle.

COMPORTEMENT ATTENDU DU MODÈLE Quelle que soit la définition, un modèle ne peut pas être réduit à sa formulation et/ou implémentation, comme le comportement attendu ou l'utilisation du modèle peuvent être vu comme des parties du modèle lui-même. Dans le cadre du perspectivisme de GIERE [GIERE, 2010c], la définition du modèle inclut le motif de l'utilisation mais aussi l'agent qui vise à l'utiliser. Pour cela une explication minimale du comportement du modèle et une exploration du rôle des paramètres est fortement recommandé pour décroître les chances de mauvais usage ou mauvaises interprétations de celui-ci. Cela inclut des graphe simple obtenus immédiatement à l'exécution sur la plateforme NetLogo, mais aussi un calcul d'indicateurs pour évaluer les sorties du modèle. Il peut aussi s'agir de visualisations améliorée pendant l'execution et l'exploration du modèle, comme le montre la figure 6.

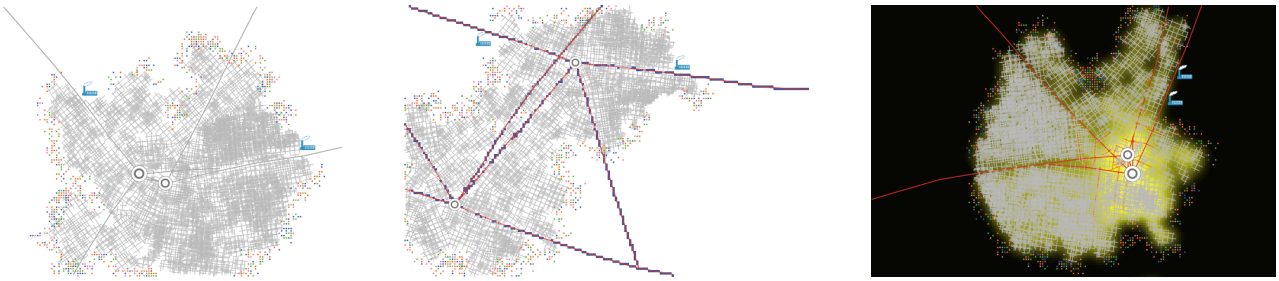


FIGURE 6 : **Exemple d'amélioration simple dans la visualisation qui peut aider à appréhender les mécanismes impliqués par le modèle.** (Gauche) Exemple de sortie originale ; (Centre) Visualisation des routes principales (en rouge) et de l'attribution des patches sous-jacente, qui suggère de possibles biais d'implémentation dans l'utilisation de la trace discrete des routes pour garder trace de leur position ; (Droite) Visualisation des valeurs foncières en utilisant un gradient de couleur plus lisible. Cette étape confirme l'hypothèse, par la forme de la distribution des valeurs, que l'étape de morphogenèse est un détour non-nécessaire pour générer un champ aléatoire pour lequel des simples mécanismes de diffusion devrait fournir des résultats similaires, comme détaillé dans le paragraphe sur l'implémentation. Initialement, l'interface du modèle ne permet pas ces options de visualisation, ces à dire se limite à la première image. On ne peut se rendre compte des processus en jeu pour la morphogenèse, liés aux patches de route et au valeurs foncières se diffusant.

Sur le besoin d'exactitude dans l'implémentation du modèle

Des divergences potentielles entre la description du modèle dans un article et les processus effectivement implémentés peut avoir des

conséquences graves sur la reproductibilité finale. Le modèle de croissance du réseau routier donné dans [BARTHÉLEMY et FLAMMINI, 2008] est un exemple d'une telle discrépance. Une implémentation stricte des mécanismes du modèle produit des résultats légèrement différents de ceux présentés dans le papier, et comme le code source n'est pas fourni nous devrions tester différentes hypothèses sur des mécanismes possibles ajoutés par le programmeur (qui semble être une règle de connexion aux intersections sous un certain seuil de distance). Des leçons qui peuvent éventuellement être tirées de cet exemple, qui rejoignent partiellement mais complètent celle tirées dans l'étude de cas précédente, sont

- la nécessité de fournir le code source
- la nécessité de fournir une description de l'architecture en même temps que le code (si la description du modèle est faite dans un langage trop loin de spécification architecturales) afin d'identifier des biais possibles d'implémentation
- la nécessité de procéder à des explorations explicites du modèle et de les détailler, ce qui dans ce cas aurait permis d'identifier de possibles biais d'implémentation.

Rendre le dernier point obligatoire pourrait assurer un risque limité de falsification puisqu'il est généralement plus compliqué de falsifier des résultats d'exploration plutôt que d'explorer effectivement le modèle. On pourrait imaginer une expérience pour tester le comportement général d'un sous-ensemble de la communauté scientifique au regard de la reproductibilité, qui consisterait en l'écriture d'un faux papier de modélisation dans l'esprit de [ZILSEL, 2015], dans lesquels des résultats opposés aux résultats effectifs d'un modèle donné seraient fournis, sans fournir l'implémentation du modèle. Un premier test serait de tester l'acceptation d'un papier clairement non reproductible dans divers journaux, si possible avec un contrôle sur les éléments textuels (par exemple en utilisant ou non des "buzzwords" chers au journal). Selon les résultats, une expérience plus poussée serait de fournir l'implémentation open source mais toujours avec des résultats modifiés plus ou moins fortement, afin de tester si les reviewers essayent effectivement de reproduire les résultats quand ils demandent le code (dans des capacités de calcul limitées bien sûr, le HPC n'étant pas encore largement disponibles en sciences sociales). Notre intuition est que les résultats obtenus seraient fortement négatifs, vu les difficultés rencontrées par une exigence de discipline de reproduction indépendante lors de nombreuses relectures, même pour des revues faisant de la reproductibilité une condition *sine qua non* de la publication, les auteurs trouvant des astuces pour se dérober aux contraintes (postuler que des données de simulation ne sont pas des données, ne fournir qu'une version agrégée inutile du jeu de données utilisées, etc. ; nous reviendrons sur le rôle des données plus loin).

3.1.2 *Exploration interactive et production des résultats*

L'usage d'applications interactives pour la fouille de données a des avantages non discutables, tel qu'une familiarisation avec la structure des données par une vue d'ensemble qui serait beaucoup plus laborieuse voire impossible autrement. C'est la même idée sous-jacente qui justifie l'interactivité pour l'exploration préliminaire des modèles basé-agent intégrée à des plateformes comme Netlogo [Wilensky, 1999] ou Gamma [Cit. gamma]. C'était d'ailleurs un objectif couplé qu'avait initialement [Rey-Coyrehourcq, 2015], c'est à dire une intégration complète de l'exploration fine des modèles et de la production des graphes de sortie ainsi que leur exploration interactive. Comme le rappelle R. Reuillon (Entretien du 11/04/2017, voir **??**), la plateforme OpenMole qui devait accueillir cette couche supplémentaire était loin d'être mature à l'époque et ne l'est toujours pas aujourd'hui, puisque l'état de l'art de telles pratiques est en pleine construction et bouleversements réguliers [Holzinger, Dehmer et Jurisica, 2014]. Des difficultés au regard de la reproductibilité, qui nous concernent particulièrement ici, sont récurrentes et loin d'être résolues. En effet, il faut bien situer la position de ces outils et méthodes comme une aide cognitive préliminaire[4], mais peu souvent comme permettant la production de résultats finaux : lorsque les paramètres ou dimension se multiplient, l'export d'un graphe est bien souvent déconnecté de l'information complète ayant conduit à sa production. De la même manière, l'utilisation de notebooks intégrés tel Jupyter, permettant d'intégrer analyses et rédaction du compte-rendu, peut devenir dangereux car on peut justement revenir sur un script, tester différentes valeurs d'un paramètre, et perdre les valeurs qui avaient produit un graphe donné. L'utilisation de versioning peut être une solution partielle mais souvent lourde. Dans l'idéal, tout logiciel interactif permettant l'export de résultats devrait en même temps exporter un script ou une description exacte et utilisable permettant d'arriver exactement à ce point à partir des données brutes. La plupart des applications d'exploration interactives de données spatio-temporelles sont à ce regard relativement immatures scientifiquement, car même dans le cas où elles sont totalement honnêtes et transparentes sur les analyses présentées à l'utilisateur, ce qui n'est malheureusement pas la règle, les tâtonnements d'exploration progressive ne sont pas reproductibles et la méthode d'extraction de caractéristiques est ainsi relativement aléatoire. En poussant le raisonnement, leur utilisation révélerait plutôt l'aveu d'une faiblesse d'un manque de méthodes systématiques accompagnant la découverte de motifs dans des données spatio-temporelles complexes de manière efficace. De manière très visionnaire, Banos avait déjà mis en garde contre "les dangers de

4 que nous ne jugeons pas superficielle puisque nous les mobilisons au moins par deux fois par la suite, voir 5.1 et 5.2

la jungle" des données dans [Banos, 2001], quand il souligne très justement que l'exploration interactive doit nécessairement se doubler d'indicateurs locaux adaptés, mais surtout d'outils d'exploration automatisés et de critère d'évaluation des choix faits et des motifs découverts par l'utilisateur. On revient encore à l'idée d'une plateforme intégrée dont OpenMole pourrait être un précurseur. La combinaison des capacités cognitives humaines au traitement machine, notamment pour des problèmes de vision par ordinateur, ouvre des possibilités de découvertes inédites, encore plus via une utilisation collective comme en témoigne le Galaxy Zoo [Raddick et al., 2010]. Les résultats d'un crowdsourcing de la cognition humaine peuvent rivaliser avec les techniques automatiques les plus avancées comme le montre [Koch et Stisen, 2017] pour l'exemple de la comparaison de cartes spatiales. Ces possibilités ne doivent cependant pas être surestimées ou utilisées à mauvais escient, et les questions d'intégration efficiente homme-machine sont d'ailleurs totalement ouvertes. Dans le domaine de la visualisation de l'information géographique, [Pfaender, 2009] introduit une sémiologie spécifique visant à favoriser l'exploration de grands jeux de données hétérogènes, et l'expérimente sur une application spécifique : il s'agit d'une avancée considérable vers une plateforme intégrée et une exploration interactive saine et reproductible, les directions d'exploration répondant à des modèles basés sur les sciences cognitives.

3.1.3 *Perspectives*

Encore une fois, la reproductibilité et la transparence sont des éléments essentiels incontournables de la science contemporaine, liés aux pratiques de science ouverte et d'accès ouvert. Beaucoup d'exemples (voir un récent en économie expérimentale dans [Camerer et al., 2016]) dans diverses disciplines montrent le manque de reproductibilité des résultats des expériences, alors que celle ci doit pouvoir conduire à une falsification ou à une confirmation de ces résultats. La falsification est une pratique coûteuse car demandant un certain investissement au détriment de sa propre recherche [Chavalarias et al., 2005]. Elle pourrait ainsi être rendue plus efficiente grâce à une transparence augmentée. Des outils spécialement dédiés à une reproductibilité directe, souvent permise par l'ouverture, devraient accroître la performance globale de la science. Mais l'accès ouvert a des impacts bien plus large que la science elle-même : [Teplitskiy, Lu et Duede, 2015] montre un transfert des connaissances scientifiques accru vers la société dans le cas d'articles ouverts, notamment par des intermédiaires comme Wikipedia.

Le développement et la systématisation de standards et de bonnes pratiques, de manière conjointe sur les différentes problématiques évoquées, est une condition nécessaire à une rigueur scientifique qui

devrait être uniforme au travers de l'ensemble des disciplines existantes. Nous construisons par exemple des exemples d'outils facilitant le flot de production scientifique, ceux-ci étant détaillés en Appendice E.3. Par exemple, pour les sciences computationnelles, on a déjà évoqué les potentialités de l'utilisation de `git` qui s'étendent en fait sans contrainte de disciplines ni de types de recherche si les bonnes adaptations sont introduites. Le suivi précis de l'ensemble des étapes d'un projet, gardé en historique offrant la possibilité de revenir à n'importe laquelle à tout moment, mais aussi de travailler de façon collaborative, plus ou moins parallèlement selon les besoins en utilisant les branches, est un exemple de service fourni par cet outil. Un exemple de bonnes pratiques d'utilisation est donné par [Perez-Riverol et al., 2016]. Plus généralement, les sciences computationnelles nécessitent l'adoption de certains standards et pratiques pour assurer une bonne reproductibilité, et ceux-ci restent majoritairement à développer : [Wilson et al., 2017] donne des premières pistes. Concernant la qualité des données, de nombreux efforts sont faits pour introduire des cadres de standardisation des données : par exemple [Veiga et al., 2017] décrit un cadre conceptuel visant à guider la résolution de problème récurrent liés à la qualité des données de biodiversité (comme par exemple évaluer des mesures jugeant de l'usage possible d'un jeu de données pour un problème donné).

C : citer Romain sur le blockchain, en lien avec ce papier ? [Furlanello et al., 2017]

L'accès aux données est également un point crucial pour la reproductibilité, et sans nous y attarder car cela impliquerait des développements sur la définition, la philosophie, le droit des données etc. qui sont des sujets de recherche en eux-même, nous donnons des perspectives sur les potentiels d'une ouverture systématique des données en recherche. En géographie, les *data paper* sont une pratique inexistante, et la règle est plutôt de garder la main jalousement sur un jeu produit, capitalisant sur le fait d'être le seul à y avoir accès. Il est évident que la qualité et quantité des connaissances produites sera nécessairement plus grande si un jeu de données est publiquement ouvert, puisqu'au moins la même chose sera obtenue, et on peut s'attendre à une prise en main par d'autres domaines, d'autres méthodes, et donc à une plus grande richesse. La fermeture induira plutôt des effets négatifs, comme par exemple du temps perdu à recoder un base vectorielle donnée uniquement sous forme de carte dans un article. L'argument du temps passé comme justification à la fermeture est absurde, puisqu'au contraire, en voyant les données comme une composante à part entière de la connaissance (voir le cadre de connaissances en 9.3), le temps passé doit impliquer plus de citations, donc plus d'utilisation, ce qui passe nécessairement par l'ouverture pour des données. De même, quelle logique, sinon la même absurde de propriété des connaissances, pousse les géographes à insérer un copyright sur l'en-

semble de leurs cartes mais aussi leurs figures, jusqu'à un copyright pour un simple histogramme qui s'en serait bien passé si on avait pu l'interroger, honnête de simplicité ? Une expérience de revue induit à réellement s'inquiéter sur la valeur donnée à l'ouverture des données par les auteurs : au bout d'une dizaine d'articles, incluant des journaux affichant comme priorité et pré-requis l'ouverture totale des données et modèles, dont un seul est seulement partiellement ouvert et l'ensemble des autres implique de croire sur parole les résultats présentés (alors qu'un des but de la revue est de contourner les biais cognitifs qu'un ou des humains ont forcément par une validation croisée qui doit se faire sur les résultats bruts et non des interprétations contenant ces biais), il est difficile de croire que des mutations profondes des pratiques ne sont pas nécessaire. Mais en suivant l'adage de Framasoft, "la route est longue mais la voie est libre", les perspectives sont nombreuses pour une évolution dont la lenteur n'est pas inéluctable. Le journal Cybergéo, pionnier des pratiques d'ouverture en sciences sociales (première revue entièrement électronique, première revue à lancer une rubrique de *model papers*), lance en 2017 une rubrique *data papers* visant à inciter le développement du partage de données et de l'ouverture en géographie. Il reste des zones grises sur lesquelles il est impossible aujourd'hui d'avoir des perspectives, notamment le droit des données. On peut citer des exemples parmi les études empiriques que nous développons : les données bibliographiques sont obtenues au prix d'une guerre de blocage par Google et un effort considérable pour la gagner ; les données immobilières proviennent d'une base propriétaire achetée avec de l'argent public, et nous pouvons profiter d'un flou du contrat pour les rendre disponibles de manière agrégées avec les résultats ; les données des stations essence proviennent d'une source dont la légalité ne devrait pas être creusée plus, et nous ne pouvons malheureusement pas les rendre disponibles sans prendre de risques - cet aspect n'a cependant jamais fait broncher les reviewers qui n'ont même pas mentionné le manque d'accès aux données. L'ouverture implique un engagement qui fait résolument partie de nos positionnements. C'est la même idée qui soutient la construction de l'application `CybergeoNetworks`[5], qui couple les outils présentés en 2.2 avec d'autres approches complémentaires d'analyse de corpus, dans le but d'encourager la réflexivité scientifique, et de mettre cet outil ouvert à la disposition d'éditeurs indépendants, pour s'émanciper de la nouvelle main mise des géants de l'édition qui à la recherche d'un nouveau modèle pour sécuriser leur profits parient sur la vente de meta-contenu et de son analyse. Heureusement, la récente loi numérique en France a gagné le bras de fer contre leur revendication d'un droit exclusif sur la fouille de texte complets.

5 `http://shiny.parisgeo.cnrs.fr/CybergeoNetworks`

⋆ ⋆

⋆

3.2 DONNÉES MASSIVES, CALCUL INTENSIF ET EXPLORATION DES MODÈLES

Nous nous positionnons à présent sur les questions liées à l'utilisation des données massives et du calcul intensif, ce qui induit par extension une réflexion sur les méthodes d'exploration de modèles. Il n'est pas évident que ces nouvelles possibilités soient nécessairement accompagnées de mutations épistémologiques profondes, et nous montrons au contraire que leur utilisation nécessite plus que jamais un dialogue avec la théorie. Implicitement, cette position préfigure le cadre épistémologique pour l'étude des Systèmes Complexes dont nous donnons le contexte à la section suivante 3.3 et que nous formalisons en ouverture 9.3.

3.2.1 *Pour un usage raisonné des données massives et de la computation*

La soi-disante *révolution des données massives* réside autant dans la disponibilité de grands jeux de données de nouveaux types variés, que dans la puissance de calcul potentielle toujours en augmentation. Même si le *tournant computationnel* ([ARTHUR, 2015]) est central pour une science consciente de la complexité et est sans douter la base des pratiques de modélisation futures en géographie comme [BANOS, 2013] souligne, nous soutenons que à la fois le *déluge de données* et les *capacités de calcul* sont dangereuses si non cadrées dans un cadre théorique et formel propre. Le premier peut biaiser les directions de recherche vers les jeux de données disponibles avec le risque de se déconnecter d'un fond théorique, tandis que le second peut occulter des résolutions analytiques préliminaires essentielles pour un usage cohérent des simulations. Nous avançons que les conditions pour la majorité des résultats dans cette thèse sont en effet ceux mis en danger par un enthousiasme inconsidéré pour les données massives, tirant la conclusion qu'un challenge majeur pour la géocomputation future est une intégration sage des nouvelles pratiques au sein du corpus existant de connaissances.

La puissance de calcul disponible semble suivre un tendance exponentielle, comme une sorte de loi de Moore. Grace à d'une part la loi de Moore effective pour le matériel, d'autre part l'amélioration des logiciels et algorithmes, conjointement avec une démocratisation de l'accès au infrastructures de simulation à grande échelle, permet à toujours plus de temps processeur d'être disponible pour le chercheur en sciences sociales (et pour le scientifique en général, mais cette mutation a déjà été opérée depuis plus longtemps dans d'autres domaines). Il y a environ une dizaine d'année, [GLEYZE, 2005] était forcé de conclure que les analyses de réseau, pour les transports publics parisiens, étaient "limitées par le calcul". Aujourd'hui la plupart des mêmes analyses seraient rapidement réglée sur un ordinateur per-

sonnel avec les logiciels et programmes appropriés : [Lagesse, 2015] est un témoin d'un tel progrès, introduisant des nouveaux indicateurs avec une plus grande complexité de calcul, qui sont calculés sur des réseaux à grande échelle. Le même parallèle peut être fait pour les modèles Simpop : les premiers modèles Simpop au début du millénaire [Sanders et al., 1997] étaient "calibrés" à la main, tandis que [Cottineau et al., 2015a] calibre le modèle Marius en multi-modélisation et [Schmitt et al., 2014] calibre très précisément le modèle SimpopLocal, chacun sur la grille avec des milliards de simulations. Un dernier exemple, le champ de la *Space Syntax*, a témoigné d'une longue route et de progrès considérables depuis ses origines théoriques [Hillier et Hanson, 1989] jusqu'à ses récentes applications à grande échelle [Hillier, 2016].

Concernant les nouvelles données "massives" qui sont disponibles, il est clair que des quantités toujours plus grandes et des types toujours nouveaux sont disponibles. De nombreux exemples de champs d'application peuvent être donnés. La mobilité en est typique, puisque étudiée selon divers points de vue, comme les nouvelles données issues des systèmes de transport intelligents [O'brien, Cheshire et Batty, 2014], des réseaux sociaux [Frank et al., 2014], ou des données plus exotiques comme des données de téléphonie mobile [De Nadai et al., 2016]. Dans un autre esprit, l'ouverture de jeux de données "classiques" (comme les applications synthétiques urbaines, les initiatives gouvernementales pour les données ouvertes) devrait pouvoir toujours plus de méta-analyses. De nouvelles façon de pratiquer la recherche et produire des données sont également en train d'émerger, vers des initiatives plus interactives et venant de l'utilisateur. Ainsi, [Cottineau, 2016] décrit une application web ayant pour but de présenter une méta-analyse de la loi de Zipf sur de nombreux jeux de données, mais en particulier inclut une option de dépôt, à travers laquelle l'utilisateur peut télécharger sont propre jeu de données et l'inclure dans la méta-analyse. D'autres applications permettent l'exploration interactive de la littérature scientifique pour une meilleure connaissance d'un horizon scientifique complexe, comme [Chasset et al., 2016] fait.

Comme toujours la situation n'est naturellement pas aussi idyllique qu'elle semble être au premier abord, et l'herbe verte du pré du voisin que nous pouvons être tentés d'aller brouter se transforme rapidement en un triste fumier. En effet, les objectifs et motivations sont flous et on peut facilement s'y perdre. Des illustrations parleront d'elles-même. [Barthelemy et al., 2013] introduit un nouveau jeu de données et des méthodes relativement nouvelles pour quantifier l'évolution du réseau de rues, mais les résultats, sur lesquels les auteurs semblent s'étonner, sont qu'une transition a eu lieu à Paris à l'époque d'Haussmann. Tout historien de l'urbanisme s'interrogerait sur le but exact de l'étude, puisque à la fin un sentiment étrange de réinven-

tion de la roue flotte dans l'air. L'utilisation des ressources de calcul peut également être exagéré, et dans le cas de la modélisation multi-agent, on peut citer [Axtell, 2016], pour lequel l'objectif de simuler le système à l'échelle 1 :1 semble être loin des motivations et justifications originelles de la modélisation agent, et pourrait même donner des arguments aux économistes *mainstream* qui dénigrent facilement les ABMS. D'autres anecdotes peuvent inquiéter : il existe en ligne des exemples étonnants, comme une application web[6] qui utilise des ressources de calcul financées par l'argent public pour simuler des distributions Gaussiennes afin de calculer pour un modèle de Gibrat, afin de calculer leur moyenne et variance, qui sont des paramètres d'entrée du modèle. En résumé, cela revient à vérifier le Théorème de la Limite Centrale. D'autre part, la distribution complète donnée par un modèle de Gibrat est entièrement connue théoriquement comme résolu e.g. par [Gabaix, 1999]. Sur ce point, nous devons partiellement être en désaccord avec le neuvième commandement de Banos, qui rappelle que "les mathématiques ne sont pas le language universel des modèles", ou plutôt souligner les dangers d'une mauvaise interprétation de ce principe[7] : il postule que des moyens alternatifs aux mathématiques existent pour faire comprendre des processus ou des méthodes, mais précise que ceux-ci sont une porte d'entrée et ne prétend jamais qu'il est possible de se passer des mathématiques, dérive que l'exemple précédent illustre parfaitement. D'ailleurs, il est possible d'exhiber des structures mathématiques très simples, comme un simplexe en dimension quelconque, dont la visualisation "simple" est un problème ouvert. Les données fournissent aussi leur collection de dérives. Récemment, sur la liste de diffusion de géographie francophone *Geotamtam*, un soudain engouement autour des données issues de *Pokemon Go* a semblé répondre plus à un besoin urgent et inexpliqué d'exploiter cette source de données avant tous les autres, plutôt qu'à des considérations théoriques élaborées. Des jeux de données existant et précis, comme la population historiques des villes (pour la France la base Pumain-INED par exemple), sont loin d'être entièrement exploités et il pourrait être plus pertinent de se concentrer sur ces jeux de données classiques qui existent déjà. De même, il faut être conscient des possibles applications de résultats basée sur des malentendus : [Louail et al., 2016] analyse la redistribution potentielle des transactions de carte bancaire au sein d'une ville, mais présente les résultats comme la base possible de recommandations de politiques pour une équité sociale en agissant sur la mobilité, oubliant que la forme et les fonctions urbaines sont couplés de manière complexe et que déplacer des transactions d'un endroit à un autre implique des

6 voir http ://shiny.parisgeo.cnrs.fr/gibratsim/

7 De manière générale, les commandements de Banos paraissent simples dans leur formulation, mais sont d'une profondeur et d'une complexité déconcertante lorsqu'on essaye d'en tirer les implications et la philosophie globale sous-jacente, et ne doivent jamais être pris à la légère.

processus bien plus complexes que des régulations directes, qui d'autant plus ne s'appliquent jamais de la façon prévue et conduisent à des résultats un peu différents. Une telle attitude, souvent observée de la part de physiciens, est très bien mise en allégorie par la figure 7 qui n'est qu'à moitié une exagération de certaines situations.

Notre principal argument est que le tournant computationnel et les pratiques de simulation seront centrales en géographie, mais peuvent également être dangereux, pour les raisons illustrées ci-dessus, i.e. que le déluge de données peut imposer les sujets de recherche et occulter la théorie, et que la computation peut éluder la construction et la résolution de modèles. Un lien plus fort est nécessaire entre les pratiques de calcul, l'informatique, les mathématiques, les statistiques et la géographie théorique. La Géographie Théorique et Quantitative est au centre de cette dynamique, puisqu'il s'agit de sa motivation initiale principale qui semble oubliée dans certains cas. Cela implique un besoin de recherche de théorie élaborées intégrées avec des pratiques de simulation conscientes. En d'autres mots, on peut répondre à des questions naïves complémentaires qui ont toutefois besoin d'être traitées une bonne fois pour toutes. Si une géographie quantitative libérée de la théorie serait possible, la réponse est naturellement non puisque cela se rapproche du piège de la fouille de données par boîte noire. Quoi qu'il soit fait par cette approche, les résultats auront un pouvoir explicatif très faible, puisqu'ils pourront mettre en valeur des relations mais pas reconstruire des processus. D'autre part, la possibilité d'une géographie quantitative purement basée sur le calcul est une vision dangereuse : même le gain de trois ordres de grandeur dans la puissance de calcul disponible ne résout pas le sort de la dimension. Prenons l'exemple des résultats de non-stationnarité obtenus en 4.1. L'utilisation de données relativement massives, de par les algorithmes spécialement conçus pour être capable de faire les traitements, est une condition nécessaire au résultat obtenus, mais à la fois l'échelle est les objets (c'est à dire les indicateurs calculés) sont co-déterminés par les constructions théoriques et les autres études empiriques. En effet l'absence de théorie impliquerait de ne pas connaitre les objets, mesures et propriétés à étudier (e.g. le caractère multi-scalaire ou dynamique des processus), et sans résolutions analytiques, il serait souvent difficile de tirer des conclusions à partir des analyses empiriques seules concernant l'ergodicité par exemple. Rien n'est vraiment nouveau ici mais cette position doit être affirmée et tenue, précisément car notre travail se base sur ce type d'outils, essayant d'avancer sur une arête fine et fragile, avec d'un côté le vide du charlatanisme théorique infondé et de l'autre l'abîme de l'overdose technocratique dans des quantités de données folles. Plus que jamais on a besoin de théories simples mais fondées et puissantes à-la-Occam [Batty, 2016], pour permettre une intégration saine des nouvelles techniques au sein des connaissances existantes.

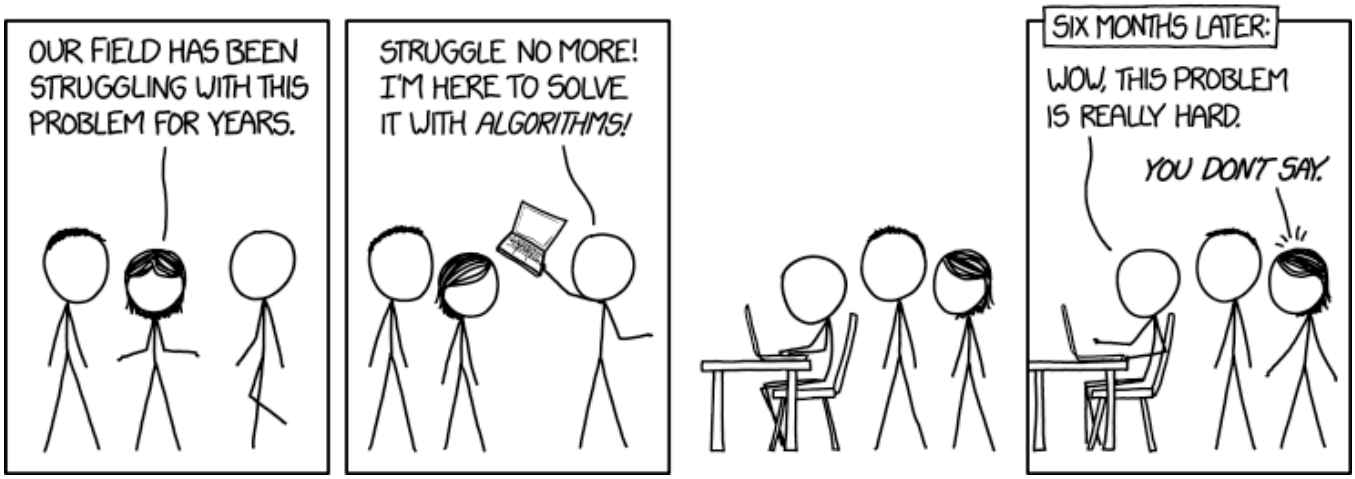


FIGURE 7 : De l'usage naïf de la fouille de données et du calcul intensif. Source : `xkcd`

3.2.2 *Contrôle statistique pour les conditions initiales par génération de données synthétiques*

Contexte

Lors de l'évaluation de modèle basés sur les données, ou même de modèle plus simples partiellement basés sur les données impliquant une paramétrisation simplifiée, une issue inévitable est le manque de contrôle sur les "paramètres implicites du systèmes" (ce qui n'est pas une notion stricte mais doit être vu dans notre sens comme les paramètres régissant la dynamique). En effet, une statistique issue d'executions du modèle sur un nombre suffisant d'executions peut toutefois rester biaisée, au sens où il est impossible de savoir si les résultats sont dus aux processus que le modèle cherche à traduire ou à une structure présente dans les données initiale. La question méthodologique fondamentale qui nous intéressera pour la suite est d'être capable d'isoler les effets propres aux processus du modèles de ceux liés à la géographie.

RATIONELLE Bien que les modèles de simulation des systèmes géographiques en général et les modèles basés-agent en particulier représentent une opportunité considérable d'explorer les comportements socio-spatiaux et de tester une variété de scenarios pour les politiques publiques, la validité des modèles génératifs est incertaine tant que la robustesse des résultats n'a pas été établie. Les analyses de sensibilité incluent généralement l'analyse des effets de la stochasticité sur la variabilité des résultats, ainsi que les effets de variations locales des paramètres. Cependant, les conditions spatiales initiales sont généralement prise pour données dans les modèles géographiques, laissant ainsi totalement inexploré l'effet des motifs spa-

tiaux sur les interactions des agents et sur leur interaction avec l'environnement. Dans cette partie, nous présentons une méthode pour établir l'effet des conditions spatiales initiales sur les modèles de simulation, utilisant un générateur systématique contrôlé par des meta-paramètres pour créer des grilles de densité utilisées dans les modèles de simulation spatiaux. Nous montrons, avec l'exemple d'un modèle agent très classique (le modèle Sugarscape d'extraction de ressources) que l'effet de l'espace dans les simulations est significatifs, et parfois plus grand que l'effet des paramètres eux-mêmes. Nous y arrivons en utilisant le calcul haute performance en un workflow très simple et open source. Les bénéfices de notre approche sont variés mais incluent par exemple la connaissance du comportement du modèle dans un contexte plus large, la possibilité de contrôle statistique pour régresser les sorties du modèles, ou une exploration plus fine des dérivées du modèle que par rapport à une approche directe.

FORMALISATION Commençons par donner une formulation abstraite de l'idée, d'un point de vue du couplage de modèle. Le générateur est considéré comme un modèle amont, couplé simplement (les sorties devenant les entrées) avec le modèle aval étudié. Si M_u est le model amont, M_d le modèle aval et α les meta-paramètres, on a la composition de la dérivée le long des meta-paramètres

$$\partial_\alpha [M_u \circ M_d] = (\partial_\alpha M_u \circ M_d) \cdot \partial_\alpha M_d$$

Cela implique que la sensibilité du modèle aval aux meta-paramètres peut être déterminée en étudiant le couplage séquentiel et le modèle amont. Nous gagnons de la connaissance thématique, dans la sensibilité à un meta-paramètre implicite, mais il y a aussi un gain computationnel : la génération de différentielles contrôlées dans l'espace initial (c'est à dire tester directement la comparaison entre deux grilles proches) serait compliquer à atteindre directement. La question de la stochasticité dans de tels modèles couplés simplement ne pose pas de problème supplémentaire puisque $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$. Cela multiplie naturellement le nombre de répétitions pour converger bien évidemment. Nous resterons dans l'application pratique ici à une étude de l'espace faisable de sortie et non à une étude différentielle, cette considération théorique n'influe pas à cet ordre, mais doit être gardée à l'esprit pour d'éventuelles applications plus fines.

ROLE DE LA DÉPENDANCE AU CHEMIN SPATIO-TEMPORELLE La dépendance au chemin spatio-temporelle est une des raisons principales rendant notre approche pertinente. En effet, un aspect crucial de la plupart des systèmes complexes spatio-temporels est leur non-ergodicité [PUMAIN, 2012b] (la propriété que les échantillons cross-sectionnels dans l'espace ne sont pas équivalent aux échantillons dans le temps pour calculer des statistiques comme la moyenne), qui témoigne généralement de forte dépendances au chemin spatio-temporelles

dans les trajectoires. De manière similaire à ce que GELL-MANN appelle *frozen accidents* dans tout système complexe [GELL-MANN, 1995], une configuration donnée contient des indices sur les bifurcations passées, qui peuvent avoir eu des effets considérables sur l'état du système. Les effets temporels et cumulatifs ont été considérés dans de nombreux sous-champs géographiques et à différentes échelles géographiques, par exemple les systèmes régionaux [WILSON, 1981] ou l'échelle intra-urbaine [ALLEN et SANGLIER, 1979]. L'impact de la configuration spatiale sur les dynamiques du modèle et les bifurcations spatiales a été moins étudié.

L'exemple des réseaux de transport est une bonne illustration, car leur forme spatiale et leur hiérarchie est fortement influencée par les décisions d'investissement du passé, les choix techniques, ou des décisions politiques qui ne sont parfois pas rationelles [ZEMBRI, 2010]. Certains indicateurs agrégés ne prendront pas en compte les positions et trajectoires de chaque agent (comme les inégalités totales dans le modèle Sugarscape) mais d'autres, comme dans le cas des motifs d'accessibilité spatiale dans un système de villes, capture entièrement la dépendance au chemin et peuvent ainsi être fortement dépendants à la configuration spatiale initiale. Il n'est pas clair par exemple ce qui a causé la transition de la capitale française de Lyon à Paris dans le bas Moyen-Age, certaines hypothèses étant la reconfiguration des motifs commerciaux du Sud au Nord de l'Europe et donc une centralité accrue pour Paris due à sa position spatiale, tout en gardant à l'esprit que les centralité géographique et politique ne sont pas équivalentes et entretiennent une relation complexe [GUENÉE, 1968]. La bifurcation induite par des facteurs socio-économiques et politiques a pris une signification profonde avec des répercussions mondiales encore aujourd'hui quand elle a été concrétisée par la configuration spatiale.

TRAVAUX EXISTANTS L'effet de la configuration spatiale sur les attributs agrégés à la zone des comportements humains a été largement discuté en géostatistiques, approximativement depuis l'introduction du *Modifiable Areal Unit Problem* (MAUP) [OPENSHAW, 1984]. Plus récemment, [KWAN, 2012] plaide pour un examen plus attentif de ce qui serait un *Uncertain Geographic Context Problem* (UGCoP), qui est la configuration spatiale des unités géographiques même si la taille et la délimitation des zones est la même. Au contraire, le faible nombre de considérations similaires dans la littérature traitant des modèles de simulation géographiques remet en question la généralisation de leur résultats, comme cela a été montré par exemple dans le cas des modèles LUTI [THOMAS et al., 2017], ou des processus de diffusion étudiés par modèles basé-agents [LE TEXIER et CARUSO, 2017].

Méthodes

Nous détaillons à présent la méthode développée pour analyser la sensibilité des modèles de simulation aux conditions spatiales initiales. S'ajoutant au protocole usuel, qui consiste à simuler un modèle μ pour différentes valeurs de ses paramètres et faire le lien entre ces variations aux variations des résultats de simulation, nous introduisons ici un générateur spatial, qui est lui-même déterminé par des paramètres et produit des ensembles de configurations spatiales initiales. Les configurations spatiales initiales sont catégorisées pour représenter des types d'espace typiques (par exemple des grilles de densité monocentriques ou polycentriques), et la sensibilité du modèle est à présent testée sur les paramètres de μ mais aussi sur les paramètres spatiaux ou les types spatiaux. Cela permet à l'analyse de sensibilité de fournir des conclusions qualitatives au regard de l'influence de la distribution spatiale sur les sorties des modèles de simulation, en parallèle des variation classiques des paramètres.

GÉNÉRATEUR SPATIAL Le générateur spatial applique un modèle de morphogenèse urbaine développé et exploré en 6.2. Pour le présenter rapidement, les grilles sont générées par un processus itératif qui ajoute une quantité de population N à chaque pas de temps, l'allouant selon un attachement préférentiel caractérisé par sa force d'attraction α. The premier processus est ensuite lissé n fois par un processus de diffusion de force β. Les grilles sont donc générées aléatoirement par la combinaison des valeurs de ces quatre meta-paramètres α, β, n and N. Pour faciliter l'exploration, seule la distribution de densité est autorisée à varier plutôt que la taille de la grille, qui est fixée à un environnement carré 50x50 de population 100,000 unités.

COMPARER LES DIAGRAMMES DE PHASE Afin de tester l'influence des conditions spatiales initiales, nous avons besoin d'une méthode systématique pour comparer des diagrammes de phase. En effet, nous avons autant de diagramme de phase que de grilles spatiales, ce qui rend une comparaison visuelle qualitative non réaliste. Une solution est d'utiliser des procédures quantitatives systématiques. De nombreuses méthodes pourraient potentiellement être utilisées : par exemple, des indicators anisotropes comme la donnée de clusters et leur position dans le diagramme de phase, peuvent permettre de révéler des *meta-transitions de phase* (transition de phase dans l'espace des meta-paramètres. L'utilisation de métriques comparant des distributions spatiales, comme la *Earth Movers Distance* qui est utilisée en viion par ordinateur pour comparer des distributions de probabilité [RUBNER, TOMASI et GUIBAS, 2000], ou la comparaison de matrices de transition agrégées de la dynamique associée au potentiel décrit par chaque distribution, est également possible. Les méthodes de comparaison de cartes, répandues en sciences environnementales,

fournissent de nombreux outils pour comparer des champs en deux dimensions [Visser et De Nijs, 2006]. Pour comparer un champ spatial évoluant dans le temps, des méthodes élaborées comme les Fonctions Orthogonales Empiriques qui isolent les variations temporelles des variations spatiales, seraient applicables dans notre cas en prenant le temps comme une dimension de paramètre, mais celles-ci ont été montrées ayant une performance similaire à la comparaison visuelle directe lorsqu'on prend la moyenne sur un ensemble de contributions crowdsourcées [Koch et Stisen, 2017]. Pour rester simple et car de telles considérations méthodologiques sont auxiliaire pour le propos principal de cette partie, nous proposons une mesure intuitive correspondant à la part de la variabilité inter-diagrammes relativement à leur variabilité interne. Plus formellement, cette distance est donnée par

$$d_r(\alpha_1, \alpha_2) = 2 \cdot \frac{d(f_{\vec{\alpha_1}}, f_{\vec{\alpha_2}})^2}{Var\left[f_{\vec{\alpha_1}}\right] + Var\left[f_{\vec{\alpha_2}}\right]} \tag{1}$$

où $\alpha \mapsto [\vec{x} \mapsto f_{\vec{\alpha}}(\vec{x})]$ est l'opérateur donnant les diagrammes de phase avec $\vec{x}$ paramètres et $\vec{\alpha}$ meta-paramètres, et d une distance entre distributions de probabilité qui peut être prise par exemple comme la distance L2 basique ou la *Earth Movers Distance*. Pour chaque valeur $\vec{\alpha_i}$, le diagramme de phase est vue comme un champ spatial aléatoire, ce qui facilite la définition des variances et de la distance.

Résultats

Sugarscape est un modèle d'extraction de ressources qui simule la distribution inégale des richesses dans une population hétérogène [Epstein et Axtell, 1996]. Des agents ayant différentes portées de vision et différents métabolismes collectent une ressource qui se régénère automatiquement et disponible de manière hétérogène dans le paysage initial. Ceux-ci s'établissent et collectent la ressource, ce qui mène certains d'entre eux à survivre et d'autres à périr. Les paramètres principaux du modèle sont le nombre d'agents, leur ressources minimale et maximale. Nous nous intéressons en prime à tester l'impact de la distribution spatiale, en utilisant le générateur spatial. La sortie du modèle est mesurée comme le diagramme de phase d'un index d'inégalité pour la distribution de la ressource (index de Gini). Nous étendons l'implémentation ayant initialement une distribution de richesse des agents, donnée par [Li et Wilensky, 2009].

Pour l'exploration, 2,500,000 simulations (1000 points de paramètres x 50 grilles de densité x 50 réplications) nous permettent de montrer que le modèle est bien plus sensible à l'espace qu'à ses autres paramètres, à la fois quantitativement et qualitativement : l'amplitude des variations entre les grilles de densité est plus grande que l'amplitude dans chaque diagramme de phase, et le comportement de ces

diagrammes de phase est qualitativement différents dans diverses régions de l'espace morphologique. Plus précisément, nous explorons une grille d'un espace de paramètre basique du modèle, dont les trois dimensions sont la population des agents $P \in [10; 510]$, la ressource minimale initiale par agent $s_- \in [10; 100]$ et la ressource initiale maximale par agent $s_+ \in [110; 200]$. Chaque paramètre est discrétisé en 10 valeurs, donnant 1000 points de paramètres. Nous procédons à 50 répétitions pour chaque configuration, ce qui donne des propriétés de convergence raisonnables. La distribution spatiale initiale varie parmi 50 grilles initiales, générée en échantillonnant les meta-paramètres du générateur dans un Hypercube Latin. Nous démontrons ainsi la flexibilité de notre cadre, par le couplage séquentiel direct du générateur avec le modèle. Nous mesurons la distance de l'ensemble des diagrammes de phase à 3 dimensions à un diagramme de phase de référence calculé sur l'initialisation du modèle par défaut (voir Fig. 8 pour sa position morphologique au regard des grilles générées), en utilisant l'équation 1 avec la distance L2 pour assurer une interpretabilité directe. En effet, cela donne dans ce cas la distance au carré moyenne entre chaque points en correspondance des diagrammes, relative à la moyenne des variances de chaque. Pour cela, des valeurs plus grandes que 1 signifient que la variabilité inter-diagramme est plus importante que la variabilité intra-diagramme.

Nous obtenons une sensibilité très forte aux conditions initiales, puisque la distribution de la distance relative à la référence s'étend sur l'ensemble des grilles de 0.09 à 2.98, avec un médiane de 1.52 et une moyenne de 1.30. Cela signifie qu'en moyenne, le modèle est plus sensible aux meta-paramètres qu'aux paramètres, et que la variation relative peut atteindre jusqu'à un facteur 3. Nous montrons en Fig. 8 leur distribution dans un espace morphologique. L'espace morphologique réduit est obtenu en calculant 4 indicateurs bruts de forme urbaine, qui sont l'index de Moran, la distance moyenne, le niveau de hiérarchie et l'entropie (voir [Le Néchet, 2015] ainsi que la section 6.2 pour une définition précise et une mise en contexte), et en réduisant la dimension avec une analyse par composantes principales pour laquelle nous gardons les deux premières composantes (92% de variance cumulée). La première mesure un "niveau d'étalement" et d'éclatement, tandis que la seconde mesure l'agrégation.[8] Nous trouvons que les grilles produisant les déviations les plus grandes sont celles avec un faible niveau d'étalement et une forte agrégation. Cela est confirmé par le comportement comme fonction des meta-paramètres, puisque des fortes valeurs de α donnent aussi une forte distance. En terme de processus du modèle, cela montre que les mécanismes de congestion induisent rapidement de plus haut niveau d'inégalités.

8 nous avons $PC1 = 0.76 \cdot distance + 0.60 \cdot entropy + 0.03 \cdot moran + 0.24 \cdot slope$ et $PC2 = -0.26 \cdot distance + 0.18 \cdot entropy + 0.91 \cdot moran + 0.26 \cdot slope$.

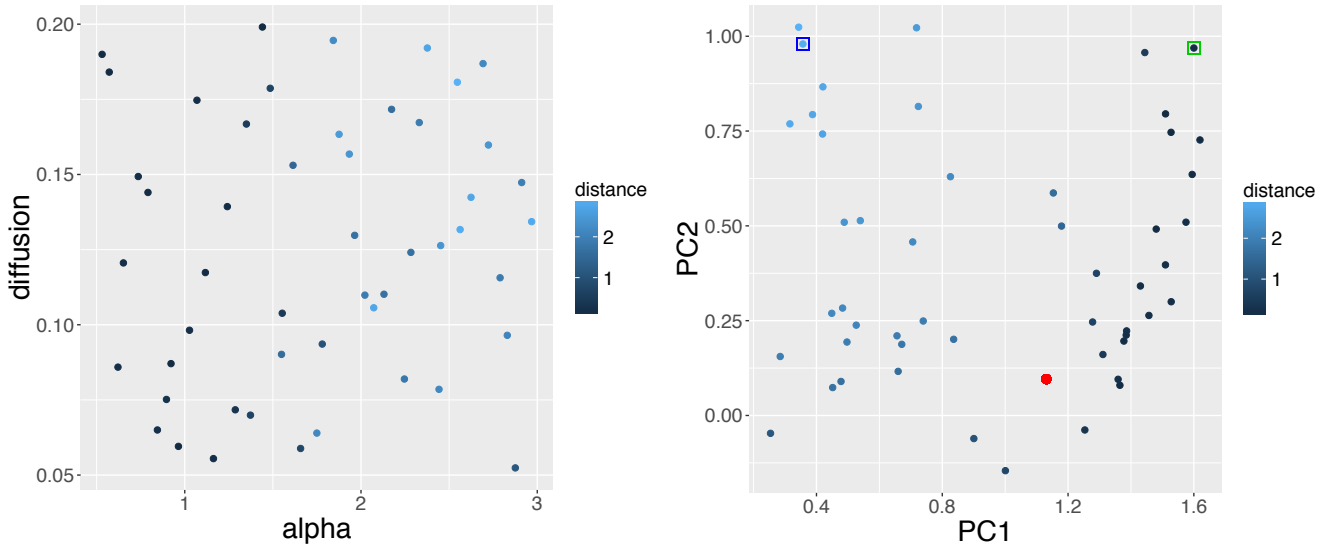


FIGURE 8 : **Distance relative des diagrammes de phase à la référence pour l'ensemble des grilles.** (Gauche) Distance relative comme fonction des meta-paramètres α (force de l'attachement préférentiel) et la diffusion (β, force du processus de diffusion). (Droite) Distance relative comme fonction des deux composantes principales de l'espace morphologique (voir texte). Le point rouge correspond à la configuration spatiale de référence. Les cadres verts et bleu donnent respectivement le premier et le second diagrammes particuliers montrés à la Fig. 9.

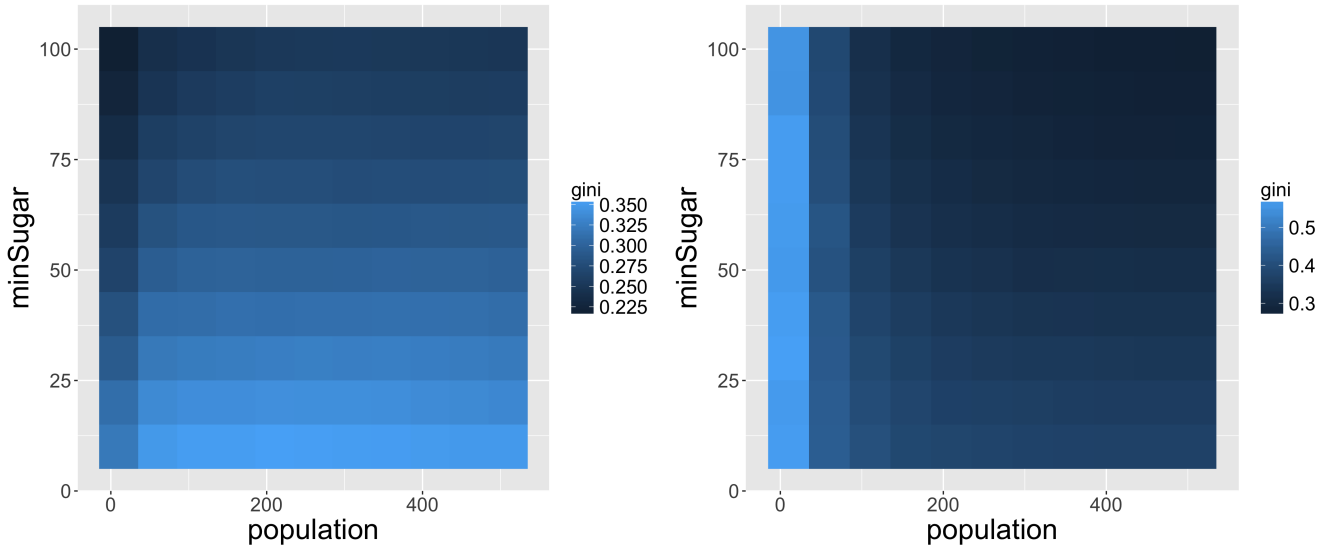


FIGURE 9 : **Exemples de diagrammes de phase.** Nous montrons deux diagrammes bi-dimensionnels sur (P, s_-), obtenus à $s_+ = 110$ fixé. (Gauche) Cadre vert, obtenu avec $\alpha = 0.79$, $n = 2$, $\beta = 0.14$, $N = 157$; (Droite) Cadre bleu, obtenu avec $\alpha = 2.56$, $n = 3$, $\beta = 0.13$, $N = 128$.

Nous contrôlons à présent la sensibilité en terme de comportement qualitatif des diagrammes de phase. Nous montrons en Fig. 9 les diagrammes pour deux morphologies très opposées en terme d'étalement, mais en contrôlant l'agrégation par la même valeur de PC2. Ceux-ci correspondent au cadres vert et bleu en Fig. 8. Les comportements sont relativement stables pour s_+ variant, ce qui signifie que les agents les plus pauvres ont un rôle déterminant dans les trajectoires. Les deux examples ont non seulement une inégalité de base très distance (le plafond du premier 0.35 est environ le plancher du second 0.3), mais leur comportement qualitatif est également radicalement opposé : la configuration étalée donne des inégalité qui décroissent quand la population décroît et qui décroissent quand la richesse minimale augmente, tandis que la concentrée donne des inégalités augmentant fortement quand la population décroît et aussi décroissantes avec la richesse minimale mais significativement seulement pour des grandes valeurs de population. Le processus est ainsi complètement inversé, ce qui aurait un impact déterminant si l'on essayait de schématiser des politiques à partir du modèle. Cet exemple confirme ainsi l'importance de la sensibilité des modèles de simulation aux conditions spatiales initiales.

3.2.3 *Lien entre modélisation et Science Ouverte*

Enfin, il est important de souligner brièvement les liens entre pratiques de modélisation et science ouverte, comme le lien entre reproductibilité et science ouverte souligné à la fin de 3.1. En fait, la Science Ouverte est composée d'un ensemble de pratiques sur différents points, d'où sa ventilation logique dans nos positionnements. Pour illustrer les enjeux, nous proposons de décrire l'exemple des workflows d'exploration de modèle comme une méthode de meta-analyse de sensibilité, c'est à dire un aspect de la méthodologie appliquée ci-dessus. Les idées de multi-modélisation et d'exploration intensive de modèle sont tout sauf nouvelles puisque Openshaw défendait déjà le "model-crunching" dans [Openshaw, 1983], mais leur utilisation effective commence seulement à émerger grâce à l'apparition de nouvelles méthodes et outils en même temps qu'une explosion des capacités de calcul : [Cottineau, Rey et Reuillon, 2016] plaide pour une approche renouvelée de la multi-modélisation. Le couplage de modèles comme nous faisons répond à des questions similaires. Dans cette lignée de recherche, la plateforme d'exploration de modèle OpenMole [Reuillon, Leclaire et Rey-Coyrehourcq, 2013] permet d'embarquer n'importe quel modèle comme une boîte noire, d'écrire des workflow d'exploration modulables qui utilisent des méthodologies d'exploration avancées comme des algorithmes génétiques, et de distribuer de manière transparente les calculs sur des infrastructures de calcul à grande échelle comme des clusters ou grilles de calcul.

Dans le cas précédent, l'outil du workflow est un outil puissant pour intégrer à la fois l'analyse de sensibilité et la meta-analyse de sensibilité, et permet de couplet n'importe quel générateur avec n'importe quel modèle de façon très directe tant que le modèle peut prendre sa configuration spatiale comme entrée ou dans un fichier d'entrée. D'autre part, une idée des workflow est de favoriser des constructions ouvertes et collaboratives, puisque le "marketplace" d'OpenMole, directement intégré au logiciel, permet de bénéficier directement des exemples qui auront été partagés sur le dépôt collaboratif. Cela ressemble aux plateformes de partage de modèles, qui sont nombreuses pour les modèles agents par exemple, mais dans un esprit encore plus modulaire et participatif. Ainsi, certains choix épistémologiques et méthodologiques au regard de la modélisation impliquent directement un positionnement au regard de la science ouverte : la multi-modélisation et les familles de modèles, qui vont de pair avec le couplage de modèle hétérogènes et multi-échelles, ne peuvent guère être viables sans des pratiques d'ouverture, de partage et de construction collaborative des modèle, comme le rappelle [Banos, 2013].

C : Sur la pédagogie : [Chen et Levinson, 2006] : la simulation comme outil pour apprendre aux élèves ingénieurs. Intéressant à utiliser pour l'aspect performatif, feedback des modèles sur les situations réelles / illustration des différents objectifs de chaque domaine : pourquoi et comment c'est intéressant de prendre en compte certains aspects selon les objectifs / perspectivisme appliqué : faire ce projet , l'évoquer ici.

⋆ ⋆

⋆

3.3 POSITIONNEMENT EPISTÉMOLOGIQUE

3.3.1 *Approche cognitive et Perspectivisme*

Notre positionnement épistémologique se fonde sur une approche cognitive de la science, donnée par GIERE dans [GIERE, 2010b]. L'approche se concentre sur le rôle des agents cognitifs comme porteurs et producteurs de la connaissance. Elle a été montrée opérationnelle par [GIERE, 2010a] qui étudie un modèle basé-agent de la science. Ces idées convergent avec le jeu Nobel de CHAVALARIAS [CHAVALARIAS, 2016] qui teste de manière stylisée l'équilibre entre exploration et falsification dans l'entreprise scientifique collective. Ce positionnement épistémologique a été présenté par GIERE comme *perspectivisme scientifique* [GIERE, 2010c], dont la caractéristique principale est de considérer toute entreprise scientifique comme une *perspective* dans laquelle des *agents* utilisent des *media* (modèles) pour représenter quelque chose dans un certain but. Pour concrétiser, nous pouvons le positionner sur la "check-list" du constructivisme de HACKING [HACKING, 1999], un outil pratique pour positionner une position épistémologique dans un espace simplifié à trois dimensions dans lequel les dimensions sont différents aspects sur lesquels les approches réalistes et constructivistes généralement divergent : d'abord la contingence (dépendance au chemin du processus de construction de connaissances) est nécessaire l'approche perspectiviste qui est pluraliste, deuxièmement le "degré de constructivisme" est assez haut car les agents produisent la connaissance, et enfin la stabilité des théories dépend des interactions complexes entre les agents et leur perspectives. Cela a pour ces raisons été présenté comme un chemin intermédiaire et alternatif entre le réalisme absolu et le constructivisme sceptique [BROWN, 2009]. la notion de *perspective* jouera un rôle fondamental dans le cadre développé en 9.3.

Cette approche mettant l'emphase sur l'auto-organisation, nous la voyons totalement compatible avec une vision anarchiste de la science comme défendue par FEYERABEND [FEYERABEND, 1993]. Celui-ci émet des doutes sur l'intérêt de l'anarchisme politique mais introduit l'*anarchisme scientifique*, qu'il ne faut pas comprendre comme un refus total de toute méthode "objective", mais d'une autorité et légitimité artificielle que certaines méthodes ou courants scientifique pourraient vouloir prendre. Il démontre par une analyse précise des travaux de Galilée que la plupart de ces résultats étaient basés sur des croyances et que la plupart n'étaient pas accessibles avec les outils et méthodes de l'époque, et postule qu'il devrait en être de même pour certains travaux contemporains. Il n'y a donc pas de *perspective* objectivement plus légitimes que d'autres dans la mesure de leurs validation par des faits et des pairs - et même dans ces cas la légitimité doit pouvoir être discutée, car la remise en question est un fondement de

la connaissance. Cela correspond exactement à la pluralité des perspectives que nous défendons. L'auto-organisation et l'émergence des connaissance nécessite un certain anarchisme pour échapper aux préconceptions cadrant par le haut. En effet, les positions anarchistes ont trouvé un écho très cohérent dans les différents courants de la complexité, de la cybernétique à l'auto-organisation au cours du 20ème siècle [DUDA, 2013]. Notre cadre de connaissance développé en 9.3 illustre cette émergence de la connaissance. De plus, notre volonté de réflexivité et de donner à notre travail des pistes de lecture diverse au delà de la linéarité (voir F), illustre l'application de ces principes. Les recommandations méthodologiques et positionnements donnés précédemment dans ce chapitre pourraient sonner comme totalitaires s'ils étaient assénés de manière sèche sans contexte, mais ceux-ci sont en fait tout le contraire puisqu'ils découlent d'un dynamique récente de science ouverte qui a bien émergé par le bas, partiellement conséquence de l'ouverture et de la pluralité.

3.3.2 *De la Vie à la Culture*

Le parallèle entre les systèmes sociaux et les systèmes biologiques est souvent fait, parfois de manière plus qu'imagée comme par exemple pour la théorie du *Scaling* de WEST qui applique des équations de croissance similaires à partir des lois d'échelle, avec des conclusions inverses tout de même concernant la relation entre taille et rythme de vie [BETTENCOURT et al., 2007]. Les relations d'échelle ne tiennent plus lorsqu'on essaye de les appliquer à une fourmi seule, et il faut alors l'appliquer à la fourmilière entière qui est alors l'organisme en question. En ajoutant la propriété de cognition, on confirme qu'il s'agit du niveau pertinent, puisque celle-ci possède des propriétés cognitives avancées, comme la résolution de problèmes d'optimisation spatiaux, ou la réponse rapide à une perturbation extérieure. Les organisations sociales humaines, les villes, peuvent-elles être vues comme des organismes ? BANOS file dans [BANOS, 2013] la métaphore de la *fourmilière urbaine* mais rappelle que le parallèle s'arrête assez vite. Nous allons voir cependant dans quelle mesure certains concepts de l'épistémologie de la biologie peuvent être utiles pour comprendre les systèmes sociaux que nous nous proposons d'étudier. Nous nous basons sur la contribution fondamentale de MONOD dans [MONOD, 1970], qui tente de développer les principes épistémologiques cruciaux pour l'étude du vivant. Ainsi, les organismes vivants répondent à trois propriétés essentielles qui permettent des les différencier d'autres systèmes : (i) la téléonomie, c'est à dire qu'il s'agit "d'objets doués d'un projet", projet qui se reflète dans leur structure et dans celles des artefacts qu'ils produisent[9] ; (ii) l'importance des processus mor-

9 à ne pas confondre avec la téléologie, propres aux animismes, qui consiste à prêter un projet ou un sens à l'univers

phogénétiques dans leur constitution (voir 6.1) ; (iii) la propriété de reproduction invariante de l'information définissant leur structure. MONOD esquisse de plus en conclusion des pistes pour une théorie de l'évolution culturelle. La téléonomie est essentielle dans les structures sociales, puisque toute organisation essaye de satisfaire un ensemble d'objectifs, même si en général elle n'y parviendra pas et que ceux-ci co-évolueront avec l'organisation. Un aspect divergent est cette notion de multi-objectif qui est typique des systèmes complexes socio-techniques. Ensuite, nous postulons que la notion de morphogenèse est un outil essentiel pour comprendre ces systèmes, avec une définition très proche de celle utilisée en biologie. Un travail approfondi pour donner cette définition est fait en 6.1, que nous résumerons en l'existence de processus relativement autonomes guidant la croissance du système et impliquant des relations causales circulaires entre forme et fonction qui témoignent d'une architecture émergente. Pour des systèmes sociaux, isoler le système est plus difficile et la notion de frontière sera moins stricte que pour un système biologique, mais on retrouvera bien ce lien entre forme et fonction, comme par exemple la structure d'une organisation ayant un impact sur ses fonctionnalités. Enfin, la reproduction de l'information est au coeur de l'évolution culturelle, par la transmission de la culture et la *mémétique*, la différence étant que le rapport d'échelle de temps entre la fréquence de transmission et les processus de croisement et de mutation ou d'autres processus non mémétiques de production culturelle est très faible, alors qu'elle est de plusieurs ordres de magnitude en biologie. [GABORA et STEEL, 2017] propose un modèle de réseau autocatalytique pour la cognition, qui expliquerait l'apparition de l'évolution culturelle par des processus analogues à ceux s'étant produit à l'apparition de la vie, c'est à dire une transition permettant au molécules de s'auto-entretenir et s'auto-reproduire, les représentations mentales faisant office de molécules. Cet exemple montre bien que le parallèle n'est pas toujours absurde. Mais si les processus à l'origine sont analogues, la nature de l'évolution est bien différente par la suite, comme le montre [LEEUW, LANE et READ, 2009], les critères darwiniens d'évolution n'étant pas suffisant pour expliquer l'évolution de nos sociétés organisées. il s'agit d'un degré de complexité supérieur et le rôle des flux d'information est crucial (voir le rôle de la complexité informationnelle dans la sous-section suivante). Enfin, l'un des points sur lequel il s'agit d'être attentif, est la plus grande difficulté de définir les niveaux d'émergence pour les systèmes sociaux : [ROTH, 2009] souligne le risque de tomber dans des cul-de-sac ontologiques car les niveaux ont été mal définis, et qu'il faut d'une manière générale penser au-delà de la seule dichotomie micro-macro qui est utilisée pour caricaturer les notions d'émergence faible, mais que les ontologies doivent souvent être multi-niveaux et impliquant de multiples niveaux intermédiaires.

C : [MESOUDI, 2017]

3.3.3 *Nature de la Complexité et Production de Connaissances*

Un aspect de la production de connaissance sur des Systèmes Complexes, auquel nous nous heurtons plusieurs fois ici (voir chapitre 9), et qui semble être récurrent voire inévitable, est une certaine réflexivité. Nous entendons par là à la fois une réflexivité pratique, c'est à dire la nécessité d'élever le niveau d'abstraction, comme le besoin de reconstruire de manière endogène les disciplines dans lesquelles une réflexion cherche à se positionner comme proposé en 2.2, ou de réfléchir à la nature épistémologique de la modélisation lors de l'élaboration d'un modèle comme en 9.2, mais également une réflexivité théorique en le sens que les appareils théoriques ou les concepts produits peuvent s'appliquer de manière récursive à eux-mêmes. Cette constatation pratique fait echo à des débats épistémologiques anciens questionnant la possibilité d'une connaissance objective de l'univers qui serait indépendante de notre structure cognitive, ou bien la nécessité d'une "rationalité évolutive" impliquant que notre système cognitif, produit de l'évolution, reflète les processus complexes ayant conduit à son émergence, et que toute structure de connaissance sera par conséquent réflexive[10]. Nous ne prétendons pas ici apporter une réponse à une question aussi vaste et vague telle quelle, mais proposons un lien potentiel entre cette réflexivité et la nature de la complexité.

COMPLEXITÉ ET COMPLEXITÉS Ce qui est entendu par complexité d'un système mène souvent à des malentendus car celle-ci peut être qualifiée selon différentes dimensions et visions. Nous distinguons d'une part la complexité au sens d'émergence faible et d'autonomie entre les différents niveaux d'un système, et sur laquelle différentes positions peuvent être développées comme dans [DEFFUANT et al., 2015]. Nous ne rentrerons pas dans une granularité plus fine, la vision de la complexité sociale donnant encore plus de fil à retordre au démon de Laplace, peut être par exemple comprise par une émergence plus forte, la nature des systèmes ne jouant pas de rôle dans notre reflexion. D'autre part, nous distinguons deux autres "types" de complexité, la complexité computationnelle et la complexité informationnelle, qui peuvent être vues comme des mesures de complexité, mais qui ne sont pas directement équivalentes à l'émergence, puisqu'il n'existe pas de lien systématique entre les trois. On peut par exemple imaginer utiliser un modèle de simulation, pour lequel les interactions entre agents élémentaires se traduisent par un message codé au niveau supérieur : il est alors possible en exploitant

10 Nous remercions D. Pumain d'avoir pointé cette vue alternative du problème que nous allons développer par la suite

les degré de liberté de minimiser la quantité d'information contenue dans le message (ce qui serait en pratique inutile car il y a des moyens plus simples de simuler un bruit blanc). Les différentes langues demandent des efforts cognitifs différents et compressent différemment l'information, ayant différents niveau de complexité mesurables [Febres, Jaffé et Gershenson, 2013]. De même, des artefacts architecturaux sont le résultat d'un processus d'évolution naturelle puis culturelle et peuvent témoigner plus ou moins de cette trajectoire. Ainsi, les liens entre ces trois types de complexité ne sont pas systématiques, et dépendent du type de système. Des liens épistémologiques peuvent néanmoins être introduits. Nous traitons ceux entre émergence et les deux autre complexités, étant donné que le lien entre complexité computationnelle et complexité informationnelle est assez bien compris et relève de problématiques de compression de l'information et de traitement du signal, ou encore de cryptographie.

COMPLEXITÉ COMPUTATIONNELLE ET ÉMERGENCE Le "paradoxe" du chat de Schrödinger n'en est un que si l'on prend une vision réductionniste, c'est à dire si l'on suppose que la superposition d'états peut se propager à travers les niveaux successifs et qu'il n'y aurait pas émergence, c'est à dire constitution d'un niveau supérieur autonome. Cette vision intuitive a récemment été démontrée rigoureusement par [Bolotin, 2014] qui prouve que l'acceptation de $\mathbf{P} \neq \mathbf{NP}$ implique une séparation qualitative entre le niveau quantique microscopique et le niveau d'observation macroscopique. En d'autres termes, la complexité computationnelle est suffisante pour avoir émergence. A priori, cette séparation effective des échelles n'implique pas que le niveau inférieur ne joue pas un rôle crucial, puisque [Vattay, Salahub et Csabai, 2015] prouve que les propriétés de criticalité quantiques sont typiques des molécules du vivant, sans qu'il n'y ait a priori de spécificité pour la vie dans cette détermination complexe par les échelles inférieures : [Verlinde, 2016] a introduit une nouvelle approche liant théories quantiques et relativité générale dans laquelle il est montré que la gravité est un phénomène émergent et que la dépendance au chemin dans la déformation de l'espace de base introduit un terme supplémentaire au niveau macroscopique, qui permet d'expliquer les déviations attribuées jusqu'alors à la "matière noire". Dans le sens inverse, le lien entre complexité computationnelle et émergence est mis en valeur par les questions liées à la nature de la computation [Moore et Mertens, 2011]. Des automates cellulaires, qui sont par ailleurs cruciaux pour la compréhension de divers systèmes complexes, ont été montrés Turing-complets (comme le Jeu de la Vie). Des organismes sans système nerveux central sont capables de résoudre des problèmes difficiles [Reid et al., 2016]. Ce lien fondamental avait été envisagé par Turing, puisqu'au delà de ses contributions fondamentales à l'informatique moderne, il s'était

intéressé à la morphogenèse et a tenté de produire des modèles chimiques d'explication de celle-ci [TURING, 1952] (qui étaient très loin d'effectivement de l'expliquer - elle n'est toujours pas bien comprise aujourd'hui, voir 6.1 - mais dont les contributions conceptuelles ont été fondamentales, notamment pour la notion de réaction-diffusion).

COMPLEXITÉ INFORMATIONNELLE ET ÉMERGENCE La complexité informationnelle, ou la quantité d'information contenue dans un système et la manière dont celle-ci est stockée, entretient également des liens fondamentaux avec l'émergence. L'information est équivalente à l'entropie d'un système et donc à son degré d'organisation - c'est ce qui a permis de résoudre le paradoxe apparent du Démon de Maxwell qui serait capable de diminuer l'entropie d'un système isolé et donc contredire la deuxième loi de la thermodynamique : celui-ci utilise en fait l'information sur les positions et vitesses des molécules du système, et son action compense la perte d'entropie par sa captation d'information. **C :** démon de Maxwell plus qu'une construction intellectuelle : [COTTET et al., 2017] at the quantum level Cette notion d'accroissement local de l'entropie a été étudiée largement par CHUA sous la forme du *Local Activity Principle*, qui est introduit comme un troisième principe de la thermodynamique, permettant d'expliquer par des arguments mathématiques l'auto-organisation pour une certaine classe de systèmes complexes typiquement impliquant des équations de réaction-diffusion [MAINZER et CHUA, 2013]. La manière dont l'information est stockée et compressée est essentielle pour la vie, puisque l'ADN est bien un système de stockage d'information (bien loin d'être compris complètement). La complexité culturelle implique un stockage de l'information bien plus complexe et à différents niveaux, et des flux d'information relevant fortement des deux autres types de complexité. Les flux d'information sont essentiels pour l'auto-organisation dans un système multi-agent. Les comportements collectifs de poissons ou d'oiseau sont des exemples typiques utilisés pour illustrer l'émergence et font partie des cas d'école de systèmes complexes. On commence cependant seulement à comprendre comment ces flux structurent le système, et quels sont les motifs spatiaux de transfert d'information au sein d'un *flock* par exemple : [CROSATO et al., 2017] introduit des premiers résultats empiriques avec l'entropie de transfert pour des poissons et pose les bases méthodologiques de ce type d'étude.

PRODUCTION DE CONNAISSANCES Nous avons à présent la matière suffisante pour en venir à la réflexivité. Il est possible de positionner la production de connaissances à l'intersection des interactions entre types de complexité développées ci-dessus. Tout d'abord, la connaissance telle que nous l'envisageons ne peut se passer d'une construction collective, et implique donc un encodage et une trans-

mission de l'information : il s'agit à un autre niveau de toutes les problématiques liées à la communication scientifique. La production de connaissances nécessite donc cette première interaction entre complexité computationnelle et complexité informationnelle. Le lien entre complexité informationnelle et émergence est mobilisé si on considère l'établissement de connaissances comme un processus morphogénétique. Il est montré en 6.1 que le lien entre forme et fonction est fondamental en psychologie : nous pouvons l'interpréter comme un lien entre information et sens, puisque la sémantique d'un objet cognitif ne peut se passer d'une fonction. Hofstader rappelle dans [Hofstadter, 1980] l'importance des symboles à différents niveaux pour l'émergence d'une pensée, qui consistent à un niveau intermédiaire en des signaux. Enfin, la dernière relation entre complexité computationnelle et émergence est celle qui nous permet d'affirmer qu'on s'intéresse particulièrement à une production de connaissance sur des systèmes complexes, les deux premiers pouvant s'appliquer à tout type de connaissance. Comme ces systèmes sont généralement multi-niveaux, ou présentent au moins un certain niveau de complexité computationnelle, leur connaissance se doit de la capturer, puisque même des modèles *simples* devront capturer leur complexité de manière conceptuelle et impliquer une structure conceptuelle sous-jacente complexe, même si celle-ci n'est pas explicitement explorée. Ainsi, toute connaissance complexe, ou *pensée complexe*, embrasse non seulement toutes les complexités mais aussi leur relations, dans son contenu et dans sa nature : elle doit nécessairement avoir un certain degré de réflexivité pour alors être cohérente. On peut tenter d'étendre à la réflexivité en tant que réflexion sur le positionnement disciplinaire : suivant Pumain dans [Pumain, 2005], la complexité d'une approche est également liée à la diversité des points de vue nécessaire pour la construire. Pour atteindre ce nouveau type de complexité, qui serait une dimension supplémentaire liée à la connaissance des systèmes complexes, la réflexivité doit être au coeur de la démarche. [Read, Lane et Leeuw, 2009] rappelle que l'innovation a été rendue possible quand les sociétés ont été capable de produire et diffuser de l'information sur leur propre structure, c'est à dire quand elles ont pu atteindre un certain niveau de réflexivité. La connaissance complexe serait donc le produit et le support de sa propre évolution grâce à la réflexivité qui a joué un rôle fondamental dans l'évolution du système cognitif : on pourrait ainsi suggérer de rassembler ces considérations, comme proposé par Pumain, sous une nouvelle notion épistémologique de *Rationalité Evolutive*.

⋆ ⋆

⋆

CONCLUSION DU CHAPITRE

La lecture d'un article ou d'un ouvrage est toujours bien plus éclairante lorsqu'on connait personnellement l'auteur, d'une part car on peut profiter des *private joke* et extrapoler certains développements des narrations qui se doivent synthétique (même si l'art de l'écriture est justement d'essayer de transmettre la majorité de ces éléments, l'ambiance en quelque sorte), et d'autre part car la personnalité a des implications complexes sur la manière d'appréhender la nature de la connaissance et une certaine structure a priori du monde. Pour cela, la connaissance scientifique serait très probablement moins riche si elle était produite par des machines aux capacités cognitives équivalentes, aux connaissances et experiences empiriques subjectives équivalentes et aussi diverses que celles humaines, mais qui auraient été programmées pour minimiser l'impact de leur personnalité et de leur convictions sur l'écriture et la communication (toujours en supposant qu'elles aient une certaine forme de données et fonctions plus ou moins équivalentes). Dans ces laboratoires de recherche dignes de *Blade Runner*, nous doutons que la production d'une pensée complexe serait effectivement possible, puisqu'il manquerait à ces machines justement la *Rationalité Evolutive* développée en 3.3, et nous doutons fortement que celle-ci puisse être produite du moins dans l'état des connaissances actuelles en intelligence artificielle. Le but de ce chapitre était donc "de faire connaissance" sur les points de positionnements incontournables pour l'ensemble de notre réflexion. Ceux-ci en sont d'autant plus en rien superflus car conditionnent très fortement certaines directions de recherche. Notre positionnement sur la reproductibilité développé en 3.1 impliquent certains choix de modélisation, notamment l'utilisation univoque de plateformes ouvertes, de workflow et d'implémentations ouverts ; il implique aussi un choix de données qui se doivent au maximum d'être accessibles ou rendues accessibles, et donc certains d'objets et d'ontologie, ou plutôt le non-choix de certains : nos problématiques pourraient être mobilisées sur des données d'entreprise fines tout en gardant une cohérence avec l'approche théorique et thématique (la théorie évolutive a largement mobilisé ce type d'étude comme par exemple [PAULUS, 2004]), mais la relative fermeture de ce type de données ne les rend pas utilisables dans notre démarche. Ensuite, notre positionnement sur le rôle du calcul intensif et les besoins d'exploration des modèles 3.2 est source de l'ensemble des expériences numériques et des méthodologies utilisées ou développées. Enfin, notre positionnement épistémologique 3.3 percole dans l'ensemble de notre travail, et permet de poser les premières briques pour des formalisations théoriques plus systématiques qui seront développées en Chapitre 9.